WAVEFIELD SINGULARITIES

(A Caustic Tale of Dislocation and Catastrophe)

by

Francis J. Wright. M.A.(Cantab.)

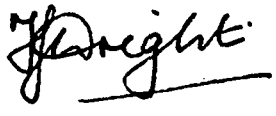
A dissertation submitted to the University of Bristol
in candidature for the degree of Doctor of Philosophy

H. H. Wills Physics Laboratory

December 1977

MEMORANDUM

This dissertation describes my own original work, except where specifically stated to the contrary. It is the result of full time research carried out between October 1974 and October 1977 in the H. H. Wills Physics Laboratory of the University of Bristol, and no part of it has previously been submitted for any degree.



SYNOPSIS

This dissertation discusses some aspects of the two principal types of singularity which can arise in wavefields: wavefront dislocations and caustics. After a general introduction in the prologue, chapter 1 develops the concept of wavefront dislocations in continuous and pulsed waves, building on the original work of Nye & Berry. Local models of scalar dislocations, and one example of a dislocation in an electromagnetic wave, are analysed. Chapter 2 examines the dislocations produced by a realistic model of an acoustic radiator by deriving an exact formula for the pulsed wavefield, and evaluating it numerically. An attempt is made in chapter 3 to derive the behaviour of the dislocations in a pulsed wavefield from the continuous wavefield.

Chapter 4 introduces the concept of caustics and the diffraction patterns which surround them. Catastrophe theory is presented in the context of caustics to show how it classifies the caustic diffraction patterns. A general method for evaluating the canonical diffraction integrals is presented, and applied in subsequent chapters. Chapter 5 summarizes the cusp diffraction pattern and draws attention to its dislocations. An ultrasonic cusp diffraction pattern is analysed, and the effect of pulsing the wavefield is considered. An appendix discusses the focussing of pulses. Chapters 6 and 7 present the three-dimensional diffraction patterns associated with the swallowtail and elliptic umbilic catastrophes, with particular emphasis on their dislocation structure. Finally chapter 8 shows that wavefront dislocations themselves are catastrophes.

To

Heather;

Tiger & Jeremy,

the animated paperweights;

and the rest of my family

for their support

and encouragement

CONTENTS

		Page
GLOSSARY		
PROLOGUE		1
CHAPTER 1	WAVEFRONT DISLOCATIONS	
1.1	Introduction	3
1.2	Strength of Wavefront Dislocations	8
1.3	Local Models for Static Dislocations	11
1.4	Static Pure Edge Dislocations	14
1.5	Lorentz Transformation of Static Dislocations	18
1.6	Local Models for Moving Dislocations	23
1.7	Moving Pure Edge Dislocations	24
1.8	Moving Screw and Mixed Dislocations	28
1.9	Dislocations in Electromagnetic Waves	30
CHAPTER 2	THE ACOUSTIC RADIATOR	
2.1	Introduction	34
2.2	Evaluation of the Geometry Function for a General Circularly Symmetric Radiator Excitation	37
2.3	Discussion of the General Formulation	39
2.4	The Rigid Circular Piston Radiator	41
2.5	The Far Field (Fraunhofer) Limit	45
2.6	Continuous Waves and their Nulls	47
2.7	Gaussian Envelope Pulse and its Dislocations	51
A2.1	Derivation of Rayleigh's Formula	57
CHAPTER 3	DISLOCATIONS BY PERTURBATION	
3.1	Introduction	59
3.2	General Perturbation Series	61
3.3	Farfield and Axial Dislocations	64
3.4	General Solution of the Perturbation Equation	70
3.5	A First Approximation	74
3.6	The Dislocation Trajectories	77
3.7	Solution of the Full Perturbation Equations	80
3.8	Continuous Wave Nulls	83
3.9	Appraisal of the Theory	85
3.10	Model Pulse with Very Simple Spectrum	87
A3.1	Decomposition of a Complex Function	90

CHAPTER 4 CAUSTICS AND CATASTROPHES

4.1 Introduction 91

4.2 Caustic Diffraction Patterns 94

4.3 Catastrophe Theory 96

4.4 The Canonical Catastrophe Diffraction Integrals 101

4.5 Integration Method 104

CHAPTER 5 THE CUSP DIFFRACTION PATTERN

5.1 The Canonical Integral and Caustic 109

5.2 Nulls by Stationary Phase 111

5.3 An Ultrasonic Cusp Diffraction Pattern 117

5.4 Dislocation Trajectories near the Cusp 121

A5.1 Computational Details 123

A5.2 Diffraction Integral for Point Sonic Source-Receiver 127

A5.3 Focussing of Pulses 130

CHAPTER 6 THE SWALLOWTAIL DIFFRACTION PATTERN

6.1 The Canonical Integral and Caustic 135

6.2 The Dark Region 137

6.3 Integration by Steepest Descent 140

6.4 The Swallowtail Diffraction Pattern and its Nulls 145

A6.1 Computational Details 150

CHAPTER 7 THE ELLIPTIC UMBILIC DIFFRACTION PATTERN

7.1 The Canonical Integral and Caustic 153

7.2 Computational Method and Results 158

7.3 Null Lines inside the Caustic 161

A7.1 Computational Details 169

CHAPTER 8 WAVEFRONT DISLOCATIONS AS CATASTROPHES

8.1 General Theory 173

8.2 The Canonical Single Strength Dislocation 177

ACKNOWLEDGEMENTS 179

REFERENCES 180

Section numbers prefixed by A are appendices

GLOSSARY

Most notation and nomenclature is standard, but some which may not be familiar is:-

$\mathrm{Re}(z)$ or z_R : real part of complex variable z

$\mathrm{Im}(z)$ or z_I : imaginary part of complex variable z

$\underline{|z}$: complex z plane

$\mathrm{int}(x)$: integral part of x (only used for $x \geqslant 0$)

$O(x)$: of order x; $y = O(x)$ if y/x is bounded in some limit

$\sim$: asymptotic equality; $y \sim x$ if $y/x = 1$ in some limit

$\theta(x)$: Heaviside unit step function; $\theta(x) = \begin{matrix} 0 \text{ if } x < 0 \\ 1 \text{ if } x > 0 \end{matrix}$

$\delta(x)$: Dirac δ-function

$\mathrm{sgn}(x)$: sign function; $\mathrm{sgn}(x) = \begin{matrix} +1 \text{ if } x > 0 \\ -1 \text{ if } x < 0 \end{matrix}$

$f'(x)$: derivative with respect to its argument; df/dx

$\mathrm{Ai}(z)$, $\mathrm{Ai}'(z)$: Airy integral function and its derivative

ret : retarded

CW : continuous wave (monochromatic)

SPP : stationary phase point

RHS/LHS : right/left hand side (of an equation)

cylinder : translation symmetric surface

localized interference fringe : line on which wave amplitude is zero

non-localized interference fringe : surface on which wave amplitude is zero (degenerate case)

head and tail of a pulse : beginning and end of the pulse with respect to time, at some fixed point in space

$\mathbb{R}^n$: real n-dimensional Euclidean space

C^∞ : class of continuous functions having infinitely many continuous derivatives

codimension : difference between dimensions of a subspace and the space in which it is embedded. Dimension of control space of a catastrophe.

diffeomorphism : differentiable 1::1 onto map whose inverse is also differentiable, e.g. an ellipse is diffeomorphic to a circle, but not to a square

dense : a set E in a space M is dense if every neighbourhood in M contains a point of E, e.g. the set of rational numbers is dense in itself, and dense in the set of real numbers, as is also the set of irrational numbers.

generic : a generic property is "not special". Then if one chooses objects at random, the chance of choosing one with a non-generic property is vanishingly small. (Rigorously, if X is a topological space which has the Baire property, that every intersection of a countable family of open dense sets is dense, then a property of elements of X is generic if the set of points of X that satisfy the property contains a set which is a countable intersection of open dense sets.)

Some important notation used mainly in chapter 1:-

β_e & β_s : complex coefficients in local model of dislocation, responsible for edge and screw character, respectively.

s : dislocation strength (an integer)

m,n : integers

ζ : z - ct

Some important notation used mainly in chapter 3:-

W : frequency deviation of pulse $(\omega - \omega_0)$

$1/\sigma$: standard deviation in time of Gaussian model pulse envelope ='n'carrier cycles

$P(\underline{r},t,W)$, $Q(\underline{r},W)$: complex factors of the wavefunction

$R(\underline{r},W)$: real factor of the wavefunction

M : amplitude of the wavefunction ,

χ, χ' : phase and frequency derivative of phase

$\overline{\chi}$: phase at t = 0; a function of space and frequency only

$Q_{(n)}(W)$: nth frequency derivative

$f^{(n)}(t)$: nth time derivative

M_x, χ_x : derivative with respect to x (essentially)

PROLOGUE

Mathematical singularities usually herald interesting physical behaviour. Wavefields exhibit two complementary types of singularity: those where the amplitude is zero, and those where the amplitude tends to infinity in the limit as the wavelength tends to zero. Their physical significance is discussed in subsequent chapters. We shall usually describe our wavefields by a complex wavefunction, so that the wave amplitude and phase are uniquely defined as the modulus and argument, respectively, of this complex wavefunction at any point in space and time. The first type of singularity is called a wavefront dislocation, for reasons discussed in chapter 1, and the singularity exhibits itself as an indeterminacy in the phase which permits the equiphase lines to coalesce into a "spider-like" pattern. The second type of singularity is called a caustic, and is only a true singularity in the "geometrical limit". It is the envelope along which the geometrical rays coalesce. In a real wavefield the amplitude is nowhere infinite, but near to the "geometrical caustic" the amplitude will attain its highest values, which is what one perceives experimentally as the caustic.

The caustics are the only readily observable feature of a wavefield: everyone has seen the cusped "teacup caustic", and the bright rippling lines on the bottom of the bath (see Berry & Nye (1977)). The recent development by René Thom and others of a branch of pure mathematics called catastrophe theory has made possible a much deeper understanding of caustics. Catastrophe theory not only classifies the possible forms of the geometrical caustics, but with the aid of a little wave theory it also classifies the diffraction patterns which "disguise" the caustics in real wavefields. This permits a general systematic study of these important regions of maximum intensity of a wave, the beginnings of which is presented here. An important feature of these caustic diffraction patterns is the wavefront dislocations which they contain, but little attention has been paid to these amplitude zeros until recently. The two types of wavefield singularity together form the backbone of any diffraction pattern, and a knowledge of their configuration provides a summary of the structure of the pattern.

However, the true significance of wavefront dislocations only becomes apparent when the wavefield is pulsed rather than continuous, because then they can, and generally do, become time dependent and move

through the wavefield. Complicated interactions among sets of moving dislocations are then possible. Their behaviour may be studied by setting up simple local models which are analytically soluble, or by setting up more realistic global models and resorting to numerical solution by computer graphics. An interesting problem is to relate the behaviour of the wavefront dislocations produced by a pulsed carrier wave to the static wavefield produced by the continuous carrier alone. We attempt to develop some general theory for this problem, and then apply it to two model systems, one of which has a caustic and one of which does not. In the latter case we can compare our predictions with exact computations of the behaviour of the dislocations, and in the former with the results of an experiment using ultrasound.

However, the straighforward way of detecting dislocations experimentally uses not their zero of amplitude, but the coalescence of a crest and trough. The latter property is probably slightly more fundamental, but less amenable to theory. Nevertheless, its consideration leads us to the fruitful conclusion that dislocations in pulses are actually catastrophes. Then we have come full circle: the pulsed diffraction patterns around caustics, which are catastrophes, contain wavefront dislocations, which are also catastrophes!

For simplicity, we consider mainly scalar wavefunctions, although all the ideas would apply to more complicated wavefunctions, such as vector and tensor waves, and quantum mechanical wavefunctions. Dislocations in continuous wave (stationary state) quantum wavefunctions have been considered by Dirac (1931), Riess (1969a,b,1975) and Hirschfelder et al. (1974,1976a,b).

CHAPTER 1

WAVEFRONT DISLOCATIONS

1.1 Introduction

The simplest possible wave is a monochromatic plane scalar wave $\psi(\underline{r},t) = A\cos(\underline{k}.\underline{r}-\omega t)$ progressing in the direction $\hat{\underline{k}}$ with wave velocity $\omega/|\underline{k}|$. At some time t,ψ has maxima and minima lying alternately in equally spaced planes perpendicular to $\hat{\underline{k}}$. We call A the amplitude of the wave and $\underline{k}.\underline{r}-\omega t = \chi(\underline{r},t)$ the phase of the wave. We call a set of planes perpendicular to $\hat{\underline{k}}$ and $2\pi/|\underline{k}|$ apart the wavefronts. These may be the planes of maxima or minima, or some intermediate set. They satisfy $\chi(\underline{r},t)$ = constant(modulo 2π), and travel along $\hat{\underline{k}}$ at the wave velocity.

If we also travel with the wave we see a fixed set of plane wavefronts analogous to a set of lattice planes in a perfect crystal. But in real crystals the lattice planes suffer from faults, among which are a set of faults localized to the neighbourhood of lines called dislocation lines. The type of a crystal dislocation line is defined by reference to the undistorted lattice planes away from the dislocation line. If the dislocation line is parallel to these planes it is called an edge dislocation and is the line along which a lattice half plane ends, as in fig 1.1. If the dislocation line is perpendicular to these planes it is called a screw dislocation and is the line along which successive planes join up in a helicoidal fashion like a screw or spiral staircase, as in fig 1.2 (see also Hilbert & Cohn-Vossen(1952), hereafter called HCV52, p209). Intermediate cases are called mixed screw-edge and are characterised by the angle of the dislocation line to the undistorted lattice planes. They also link successive lattice planes; only the pure edge dislocation does not. Similarly, the wavefronts of a more realistic wavefield may suffer from analogous faults, called "wavefront dislocations" by Nye & Berry in their pioneering paper of 1974 (hereafter called NB74). Therein they present the basic theory of wavefront dislocations, which is fundamental to the present work. In this chapter we review these basic concepts and endeavour to generalize them slightly.

The fundamental property of a dislocation is that if one traverses <u>any</u> closed circuit enclosing the dislocation line, then there exist

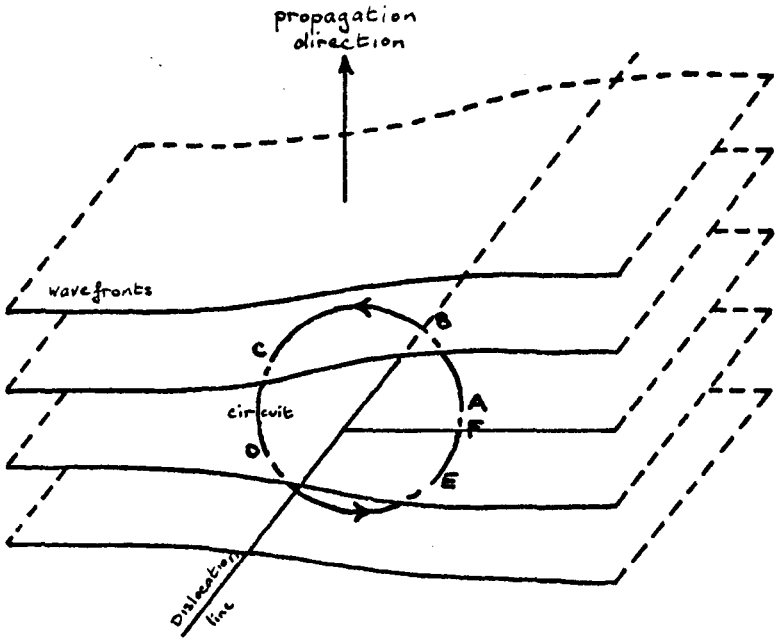


Figure 1.1 Single Edge Dislocation

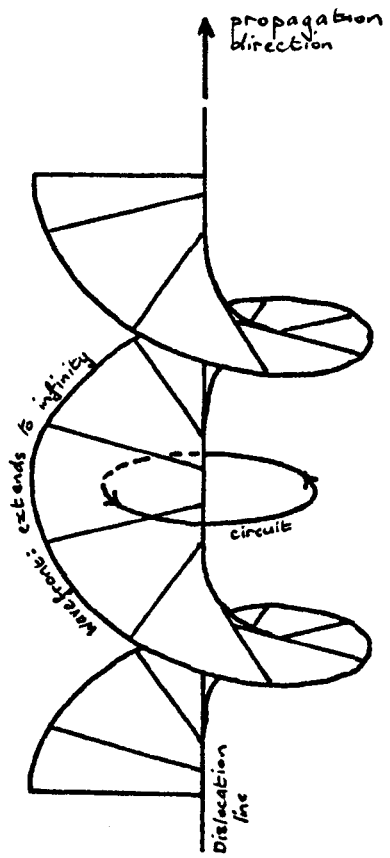


Figure 1.2 Single Screw Dislocation
(a semi-helicoid)

wavefronts through which one passes an odd number of times. The number of such wavefronts is the strength n of the dislocation. The dislocations illustrated in figs. 1.1 & 1.2 are both of strength 1, and an undislocated region can be considered as a dislocation of strength 0. The closed circuit is the analogue of the Burger's circuit for a crystal dislocation.

To show that wavefront dislocations may occur, let us consider the most general monochromatic scalar wavefunction $\psi(\underline{r},t)$ in the form

$$\psi(\underline{r},t) = \mathrm{Re}\left\{A(\underline{r})e^{i\{B(\underline{r})-\omega t\}}\right\}$$

where $B(\underline{r})$ is real. We can write the complex function $A(\underline{r})$ in modulus-argument form as

$$\rho(\underline{r})e^{i\phi(\underline{r})}$$

giving $\psi(\underline{r},t) = \rho(\underline{r})\cos\{\phi(\underline{r})+B(\underline{r})-\omega t\}$, where everything is real and $\rho(\underline{r})\geqslant 0$. Again $\rho(\underline{r})$ is the (variable) amplitude and $\chi(\underline{r},t) = \phi(\underline{r})+B(\underline{r})-\omega t$ is the phase. Now the maxima and minima in time correspond to certain values of phase exactly as in the plane wave case, but the maxima and minima in space correspond to different values of phase. Therefore we define the wavefronts to be the equiphase surfaces, satisfying $\chi(\underline{r},t)$ = constant(modulo 2π), so that the phase increases continuously through 2π between wavefronts. Since the wavefunction is continuous, the phase must vary continuously along a loop enclosing a dislocation line. Let us start at point A in fig. 1.1 with phase α and suppose the phase increases upwards. Then the phases at points B,C,D,E,F are respectively $\alpha+2\pi$, $\alpha+2\pi$, α, α, $\alpha+2\pi$. But A and F are the same point, so we see that the phase has increased by 2π in traversing the loop. This is allowed because phase is only defined up to a multiple of 2π and $\psi(\underline{r},t)$ is still single valued. The loop around the screw dislocation behaves identically, and obviously for a dislocation of strength n the phase change round the loop would be $2\pi n$.

Now let us progressively shrink the loop whilst retaining the the continuous change of phase of $2\pi n$ round the loop. When the loop has shrunk to a point, the phase at that point must have all values in a range of $2\pi n$. But $\psi(\underline{r},t)$ must be single valued. This is only possible if the amplitude $\rho(\underline{r}) = 0$. Then $\psi(\underline{r},t)$ is well defined to be zero and the phase $\chi(\underline{r},t)$ is undefined.

$$\left.\begin{aligned} &\rho(\underline{r}) = 0 \\ \Rightarrow\ &A(\underline{r}) = 0 \quad \text{(complex)} \\ \Rightarrow\ &\mathrm{Re}\ A(\underline{r}) = \mathrm{Im}\ A(\underline{r}) = 0 \end{aligned}\right\} \qquad (1.1)$$

and these two equations define a line in three dimensions. This is the equation of the dislocation line, and this definition of the dislocation line is a direct consequence of our definition of wavefronts as equiphase surfaces. (An alternative definition is mentioned in NB74 and discussed in the last chapter.) These dislocation lines are fixed in space and are just localized interference fringes - lines of zero intensity.

There is also a degenerate case where $\mathrm{Re}\ A(\underline{r}) \propto \mathrm{Im}\ A(\underline{r})$, and (1.1) reduces to a single equation which only defines a surface in three dimensions. We can write the wavefunction as $\psi(\underline{r},t) = A(\underline{r})\cos\{B(\underline{r})-\omega t\}$ where $A(\underline{r})$ is real. Generically $A(\underline{r})$ passes linearly through zero thereby changing sign on opposite sides of the null surface. On the positive side $\psi(\underline{r},t) = |A(\underline{r})|\cos\{B(\underline{r})-\omega t\}$, but on the negative side $\psi(\underline{r},t) = |A(\underline{r})|\cos\{B(\underline{r})-\omega t\pm\pi\}$. The amplitude, which is non-negative by definition, is $|A(\underline{r})|$ and the phase jumps by π on crossing the null surface. The wavefronts have the "staggered comb" structure of fig. 1.3 (or fig. 5 of NB74). This structure is a non-localized interference fringe, not really a dislocation, and normally only occurs in the far field limit of diffraction problems, such as the elementary analysis of Young's slits or the far field of the acoustic radiator (see the next chapter). In the general case where $A(\underline{r})$ is essentially complex we still expect it to pass through 0 linearly. Therefore if we travel along a smooth curve intersecting the dislocation line we still expect the phase to jump by exactly π as we cross the dislocation line. This fact is useful when plotting phase lines around a dislocation and allows us to deduce that generically the phase in a two-dimensional section through a single strength dislocation has locally the "spider" pattern of fig. 1.4 and globally the pattern of fig. 1.5 (cf. fig. 10 of NB74) showing a wavefront ending. Note the essential saddle point (here at phase π), and that two-dimensional sections through dislocations always appear to be of pure edge type.

The most interesting case is when we make the dislocations time dependent by making the amplitude a function of time $\rho(\underline{r},t)$ and the phase a more general function of time. We do this by modulating the wave source. To be able still to apply the previous theory we require that the modulation envelope varies much more slowly in time than the

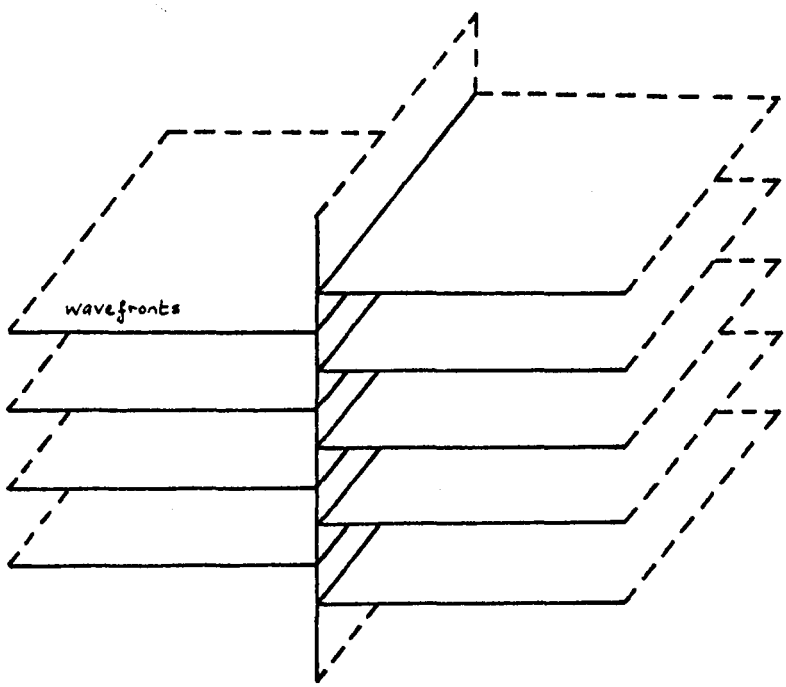


Figure 1.3 Non-localized interference fringe

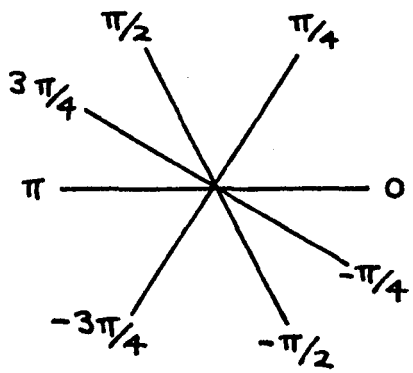


Figure 1.4 Local Phase around Dislocation

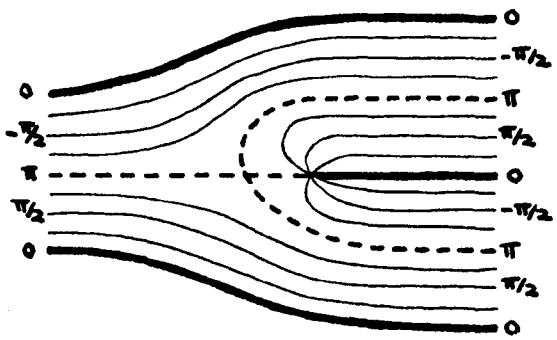


Figure 1.5 Global Phase around Dislocation

carrier. We call this a quasimonochromatic wave, and strictly the condition for this is that the width of the frequency spectrum be much less than its centre frequency. We wish to consider a wavefunction of the form $\psi(\underline{r},t) = \rho(\underline{r},t)\cos\{\phi(\underline{r},t)-\omega t\}$. But now the amplitude and phase are not uniquely defined by $\psi(\underline{r},t)$ alone, since we can change ρ arbitrarily and compensate by a change of ϕ. We need an additional piece of information such as the "complementary wavefunction" $\psi_c(\underline{r},t) = \rho(\underline{r},t)\sin\{\phi(\underline{r},t)-\omega t\}$ i.e. we need both the real and imaginary parts of the complex wavefunction

$$\psi(\underline{r},t) = \rho(\underline{r},t)e^{i\{\phi(\underline{r},t)-\omega t\}}$$

from which uniquely:

amplitude $\rho(\underline{r},t) = |\psi(\underline{r},t)|$

phase $\chi(\underline{r},t) = \phi(\underline{r},t)-\omega t = \arg\{\psi(\underline{r},t)\}$

In the monochromatic (or continuous wave) case the real wavefunction $\psi(\underline{r},t) = \rho(\underline{r})\cos\{\phi(\underline{r})-\omega t\}$ contains all the necessary information, because we can derive the complementary wavefunction from it, for example by taking a later time $t' = t+\pi/(2\omega)$ or taking the rate of change in time, viz

$$\psi(\underline{r},t') = \rho(\underline{r})\cos\{\phi(\underline{r})-\omega t-\pi/2\} = \rho(\underline{r})\sin\{\phi(\underline{r})-\omega t\}$$

or $$\frac{\partial\psi(\underline{r},t)}{\partial t} = \omega\rho(\underline{r})\sin\{\phi(\underline{r})-\omega t\}$$

so that $$\psi_c(\underline{r},t) = \psi(\underline{r},t') = \frac{1}{\omega}\frac{\partial\psi(\underline{r},t)}{\partial t}$$

(In fact, if the modulation is varying sufficiently slowly in time, i.e. the wave is sufficiently quasimonochromatic, it is possible to derive the amplitude $\rho(\underline{r},t)$ and phase $\chi(\underline{r},t)$ approximately from the real wavefunction alone. If the original carrier wave is available, of course, it can be done exactly for a quasimonochromatic wave, and this is the basis of an amplitude-phase (Argand) display produced by Walford et al.(1977).)

Note that now neither the maxima or minima in space nor in time correspond to specific values of phase, i.e. to wavefronts as defined earlier. Section 2 of NB74 discusses how using this complex wavefunction corresponds experimentally to measuring a real wavefunction, then phase shifting the carrier by $\pi/2$ and observing another real wavefunction and combining the results. The significance of $\rho(\underline{r},t)$ is that it is the envelope of the wavefunction as the phase of the source carrier is

varied. However, in practice this procedure is rarely followed and only a single real wavefunction is observed. The striking property of a dislocation is the appearance or disappearance of a crest and trough (maximum and minimum) at some point near the dislocation line as defined earlier. This approach is taken up again in the final chapter. Even if a complex wavefunction is not observed in practice, it is eminently suited to theoretical investigations, and henceforth we shall always work with the complex wavefunction

$$\psi(\underline{r},t) = \rho(\underline{r},t)e^{i\chi(\underline{r},t)}$$

such that a dislocation line has the very simple equation

$$\rho(\underline{r},t) = 0 .$$

1.2 Strength of Wavefront Dislocations

Given a complex wavefunction

$$\psi(\underline{r},t) = \rho(\underline{r},t)e^{i\chi(\underline{r},t)}$$

we wish to know

(a) where the dislocation lines are at some time t

(b) what strength they have.

If we admit the trivial case of a zero of amplitude around which there is no net phase change as a dislocation of strength 0, then a necessary and sufficient condition for a dislocation is

$$\psi(\underline{r},t) = 0 \quad (\text{or } \rho(\underline{r},t) = 0).$$

Having found the dislocation lines, if we traverse a circuit C in space-time, then the net number of wavefronts crossed, i.e. the number of units of 2π by which the phase changes, is the total strength s of the dislocations encircled by the loop:

$$s = \frac{1}{2\pi}\oint_C d\chi$$

The sign of s depends on which way we go round C. By a circuit in space-time is meant the following: a dislocation line moving with time constitutes a one-parameter family of lines in 3-space, parametrised by time. Such a family of lines constitutes (generally) a 2-surface. In 3-space this surface is that traced out as the dislocation line moves - we call it the dislocation trajectory. But we can also embed this trajectory surface in four dimensional space-time such that the real dislocation line at time t_0 is the intersection of the hyperplane $t = t_0$ with this "4-trajectory" surface, and the normal trajectory surface is the projection of the 4-trajectory onto real 3-space. But the generic intersection of <u>any</u> hypersurface with the 4-trajectory is a line, although not the physical dislocation line. We may take as our space-time circuit any circuit lying in the hypersurface and encircling this line. In view of this we will sometimes drop the distinction between space and time, and represent $(\underline{r},t)$ by the 4-vector x (having components x^μ : $\mu=1,2,3,4$).

The complex wavefunction $\psi(x)$ is a mapping from space-time into the complex (Argand) plane, which will map a closed circuit C in space-time into a closed contour C' in the complex plane (closed because $\psi(x)$ is single valued). The strength of the dislocations encircled by C is

the number of times C' encircles the origin of the complex plane. This is a very useful way of finding the strength of a model dislocation, by choosing the simplest possible space-time contour C. It also shows that the strength is a topological invariant of a dislocation. If we continuously distort C, then C' will also be continuously distorted because $\psi(x)$ is continuous. The number of times which C' encircles the origin can only change if C' crosses the origin, which means that C has crossed the dislocation line. Therefore all topologically equivalent circuits which encircle the dislocation line once lead to the same strength. Also a dislocation line cannot end, because if it did the circuit C could be shrunk to a point not on the dislocation line without passing through it. This implies that C' could be shrunk to a point other than the origin without passing through the origin, which is clearly impossible. Thus dislocation lines can only pass out of the region where a wavefield is defined (i.e. end on boundaries such as reflecting or radiating surfaces) or form closed loops. They may, however, interact with each other such that the total strength is conserved.

It should be noted that the sign of the strength of a dislocation is not absolute, but is relative to a family of topologically equivalent directed circuits. For a set of interacting dislocations one must take a circuit C which encloses the whole set, and then find the signed strength of each member of the set on its own relative to a circuit topologically equivalent to C. The total strength of the set is then the algebraic sum of these component strengths, and this must be conserved in any interactions.

An alternative approach, which we shall use later, is to always choose our circuit round a single dislocation line such that s is positive (say), and indicate the sense of this circuit by attaching a direction to the dislocation line itself, via the right hand (corkscrew) rule (say). Then for a set of interacting dislocations the total strength of the dislocations within a given circuit is the sum of the strengths passing through the circuit in one sense minus the sum of the strengths in the other sense.

It is possible to find a general expression for the strength of a dislocation without explicit reference to phase, by a method based on the complex residue calculus:

$$\psi = \rho e^{i\chi}$$

$$\Rightarrow \frac{d\psi}{\psi} = \frac{d\rho}{\rho} + i d\chi$$

$$\Rightarrow d\chi = \mathrm{Re}\left\{ \frac{1}{i} \frac{d\psi}{\psi} \right\}$$

$$\Rightarrow \text{strength } s = \mathrm{Re}\left\{ \frac{1}{2\pi i} \oint_C \frac{d\psi(x)}{\psi(x)} \right\}$$

$$= \mathrm{Re}\left\{ \frac{1}{2\pi i} \oint_{C'} \frac{d\psi}{\psi} \right\}$$

But by residue calculus

$$\frac{1}{2\pi i} \oint_{C'} \frac{d\psi}{\psi} = n$$

if C' encircles the pole (of $1/\psi$) at the origin n times, and is real. Therefore,

$$\text{strength } s = \frac{1}{2\pi i} \oint_C \frac{d\psi(x)}{\psi(x)}$$

This is, of course, just a restatement of the previous contour condition, and is not used further.

1.3 Local Models for Static Dislocations

This section is based on sections 4 & 5 of NB74, and following that work, we shall consider wavefunctions which are simply modulated plane waves travelling along the z axis. The simplest such wavefunctions travel "rigidly" along the z axis, i.e. z and t appear only in the combination $\zeta = z-ct$, where c is the wave velocity; we shall call the dislocations associated with such wavefunctions static because they do not move relative to the wavefronts. This is to be contrasted with the dislocations of monochromatic waves which do not move in space, but the wavefronts sweep through them.

We consider the plane wave $e^{ik\zeta}$ to be modulated by a "complex envelope function" $\Psi(\underline{r},t)$ of space and time. Our local models will consist of the lowest order polynomial approximations to Ψ in the neighbourhood of the dislocation line which produce the required behaviour. They must, of course, satisfy the scalar wave equation

$$\nabla^2\psi(\underline{r},t) = \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}(\underline{r},t)$$

but they will not be acceptable as global wavefunctions because the amplitude will diverge (polynomially) away from the dislocation line, and they could not therefore satisfy any realistic boundary conditions. This would require the addition of higher order terms, which would not, however, affect the behaviour near the dislocation.

Our general model dislocated wavefunction has the form

$$\psi(\underline{r},t) = \Psi(\underline{r},t)e^{ik\zeta},$$

and for static dislocations this simplifies to

$$\Psi(x,y,\zeta)e^{ik\zeta}.$$

Then the fact that $\psi(\underline{r},t)$ satisfies the wave equation implies that $\Psi(x,y,\zeta)$ satisfies the two-dimensional Laplace equation

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} = 0,$$

the general solution of which may be written as

$$\Psi(x,y,\zeta) = f(x+iy,\zeta) + g(x-iy,\zeta) \qquad (1.2)$$

where f and g are arbitrary complex functions.

The equation of the dislocation lines of $\psi(\underline{r},t)$ is $\Psi(x,y,\zeta) = 0$. Let us attempt to model a single straight dislocation line passing through the origin. Then $\Psi(x,y,\zeta)$ must take the form

(linear expression)^s

i.e. $$\Psi(x,y,\zeta) = \left[a(x+iy) + b(x-iy) + c\zeta\right]^s \qquad (1.3)$$

where a,b,c are complex coefficients. But to satisfy the wave equation $\Psi(x,y,\zeta)$ must have the form (1.2).

If s = 1 then (1.3) does have the required form, and factoring out a complex constant we can write it as

$$\Psi(x,y,\zeta) = x + \beta_s y + \beta_e \zeta \qquad (1.4)$$

where the coefficients β_s and β_e are not both real (otherwise the phase would be constant). This is a slightly generalized version of eqn.(24) of NB74. The equation of the dislocation line D is

$$\left.\begin{aligned} x + \beta_{sR} y + \beta_{eR}\zeta = 0 \\ \beta_{sI} y + \beta_{eI}\zeta = 0 \end{aligned}\right\}$$

If $\beta_e = 0$, then x=y=0, D lies along the ζ axis perpendicular to the (unmodulated) wavefronts and we have a pure screw dislocation. If β_s is real then $\zeta = x+\beta_s y = 0$, D lies in the xy plane parallel to the wavefronts and we have a pure edge dislocation (cf. eqn.(13) of NB74). Otherwise D lies at some angle to the xy plane and we have a mixed screw-edge dislocation.

We shall illustrate the contour method for finding the strength of a dislocation in this simplest case. Since

$$k\oint d\zeta = 0$$

the plane wave factor $e^{ik\zeta}$ does not affect the strength and only the envelope function Ψ need be considered. Suppose D is not pure edge, then we may take our circuit C in the xy plane around the origin (fig. 1.6). As long as β_s is not pure real, i.e. D does not lie in the plane of C, then the circuit C' encircles the origin once, and we have a single strength dislocation. So (1.4) is the envelope function of the general straight single strength dislocation.

If $s \neq 1$ in (1.3), then when we multiply out the repeated factors we shall produce cross terms of the form $(x+iy)^n(x-iy)^m$, which do not exist in the form (1.2), unless either a = 0 or b = 0. Then factoring

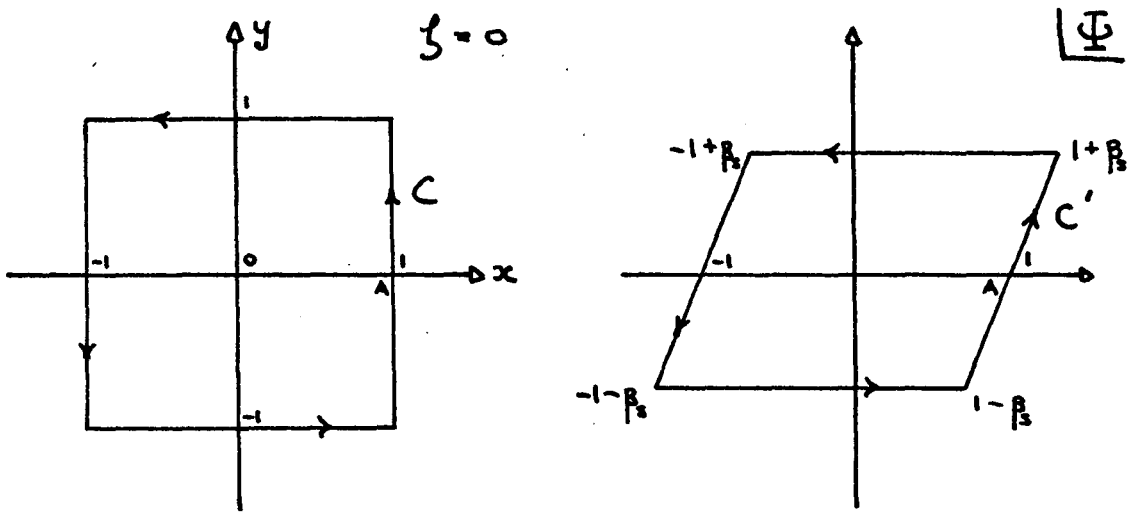


Figure 1.6 Circuits for Single Strength Dislocation (1.4)

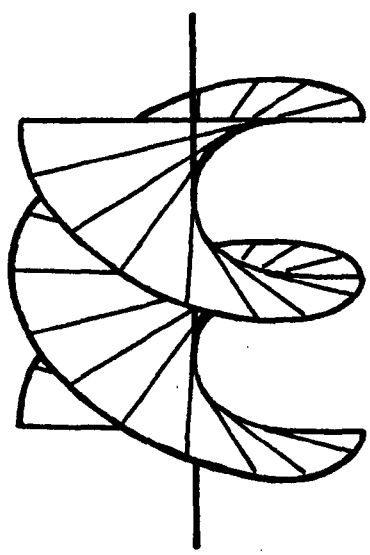


Figure 1.7 Double Screw Dislocation
(a right helicoid)

out a complex constant, we have

$$\Psi(x,y,\zeta) = \left[(x \pm iy)+\beta\zeta\right]^s \tag{1.5}$$

where β is an arbitrary complex coefficient. The equation of the dislocation line D is

$$\left.\begin{aligned} x + \beta_R\zeta &= 0 \\ \pm y + \beta_I\zeta &= 0 \end{aligned}\right\}$$

If $\beta = 0$, D is pure screw, otherwise it is mixed screw-edge. It cannot be pure edge, even in the limit $|\beta| \to \infty$, because $(\beta\zeta)^s$ would have constant phase. The projection of D onto the xy plane makes an angle $\pm\arg(\beta)$ with the x axis, and if δ is the inclination of D to the wavefronts (xy plane) then $\tan\delta = 1/|\beta|$ (cf. NB74 p180). If we make our circuit C around D then the factor $[-\bullet-]$ obviously makes one circuit around the origin, as before. The phase of Ψ is s times that of this single factor, and therefore makes s circuits. So (1.5) represents a mixed screw-edge dislocation of strength s (which exists also in the pure screw limit $\beta = 0$), contrary to the conjecture on p182 of NB74.

A multiple pure screw dislocation may be written

$$\psi(r,\phi,\zeta) = r^s e^{i\{k\zeta+s\phi\}}$$

where $x = r\cos\phi$, $y = r\sin\phi$ (cf. NB74 eqn.(30)). The equation of the wavefronts is $k\zeta + s\phi = c + 2n\pi$ where c is some constant. Thus

$$\phi = (k\zeta - c - 2n\pi)/s$$

At fixed ζ this is s radial straight lines given by $n = 0,1,2, \ldots(s-1)$, and the wavefronts are generated by "screwing" this configuration up the ζ axis. For $s = 2$ the surface is a right helicoid, which is a ruled surface of constant pitch (e.g. see fig. 219, p209 of HCV52) as in fig, 1.7. The single strength screw of fig. 1.2 is just half of this right helicoid.

To summarize, we can write the basic static straight dislocation of strength s as

$$\Psi(x,y,\zeta) = \left[x + \beta_s y + \beta_e\zeta\right]^s$$

where β_s and β_e are not both real, and $\beta_s = \pm i$ unless $s = 1$.

1.4 Static Pure Edge Dislocations

We have constructed a straight single edge dislocation, but the edge limit of the multiple mixed dislocation does not exist. Is it then possible to have multiple pure edge dislocations? NB74 conjectures that it is not, and we shall attempt to prove this.

Suppose we look at a section through a multiple mixed dislocation e.g. assume β is not pure real and consider the section $y = 0$ of (1.5):

$$\Psi(x,0,\zeta) = [x+\beta\zeta]^s$$

Now if we impose y translation symmetry on this function and construct

$$\Psi(x,y,\zeta) = [x+\beta\zeta]^s$$

we appear to have the edge dislocation line $x = \zeta = 0$ of strength s. But unfortunately it does not satisfy the wave equation, in either two or three dimensions! We now formalize this argument.

The condition that a dislocation be pure edge is highly restrictive in that the dislocation line D must lie entirely in the plane of the wavefronts, so that a static edge dislocation line must satisfy

$$\zeta = 0$$

and its equation in the xy plane will be

$$f(x,y) = 0$$

where f is an arbitrary <u>real</u> function. Then we require

$$\Psi(x,y,\zeta) = [f(x,y)+\beta\zeta^m]^s$$

with β not pure real (otherwise the phase is constant). Consider traversing a circuit C around the dislocation. Then C' will encircle the origin s times if m is odd, and f(x,y) is locally antisymmetric about its zeros, otherwise C' will double back on itself, as in fig. 1.8.

But $\Psi(x,y,\zeta)$ must also satisfy the two-dimensional Laplace equation:

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} = 0$$

$$\Rightarrow s(f+\beta\zeta^m)^{s-2}\left[(s-1)\left\{\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial x}\right)^2\right\}+(f+\beta\zeta^m)\left\{\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}\right\}\right] = 0$$

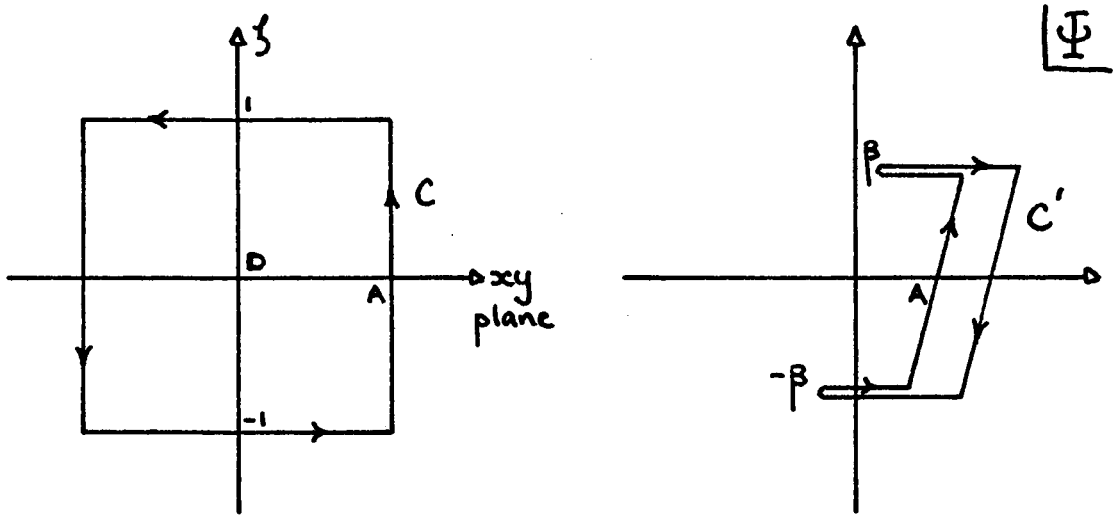


Figure 1.8 Circuit for f(x,y) Locally Symmetric (s = 1)

Now $s(f+\beta\zeta^m)^{s-2} \neq 0$ (except on a dislocation line if $s>2$), so we equate real and imaginary parts of $\left[\rightarrow\right]$ to 0.

$$\text{real: } (s-1)\left\{\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2\right\} + (f+\beta_R\zeta^m)\left\{\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}\right\} = 0$$

$$\text{imag: } \beta_I\zeta^m\left\{\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}\right\} = 0$$

$$\Rightarrow \frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2} = 0 \qquad (1.6)$$

i.e. $f(x,y)$ satisfies the two-dimensional Laplace equation and

$$(s-1)\left\{\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2\right\} = 0$$

If $s \neq 1$, then $\partial f/\partial x = \partial f/\partial y = 0$ since f is real. This implies f = constant, which satisfies (1.6) but does not produce a dislocation, so $s = 1$ is the only solution. Therefore, the only possible static pure edge dislocation has single strength, with

$$\Psi(x,y,\zeta) = f(x,y) + \beta\zeta^{2n+1} \qquad (1.7)$$

where β is not pure real, n is a non-negative integer and $f(x,y)$ is any real solution of the two-dimensional Laplace equation (1.6) which is locally antisymmetric about its zeros.

The only linear dislocation of this form is

$$\Psi(x,y,\zeta) = x + Ay + \beta\zeta^{2n+1}$$

where A is real, which corresponds to the pure edge limit of (1.4) if we take $n = 0$. If we try to generalize it to a mixed single strength dislocation by making A complex, we produce a curved dislocation line unless $n = 0$.

The next simplest example is when $f(x,y)$ is quadratic, and the most general real quadratic solution of (1.6) is

$$f(x,y) = ax^2 + 2hxy - ay^2 + 2fx + 2gy + c$$

with all coefficients real. $f(x,y) = 0$ is always the equation of a rectangular hyperbola, so by a change of coordinates we can write

$$f(x,y) = xy - c \quad \text{(a hyperbolic paraboloid)}$$

and the dislocation line is shown in fig. 1.9.

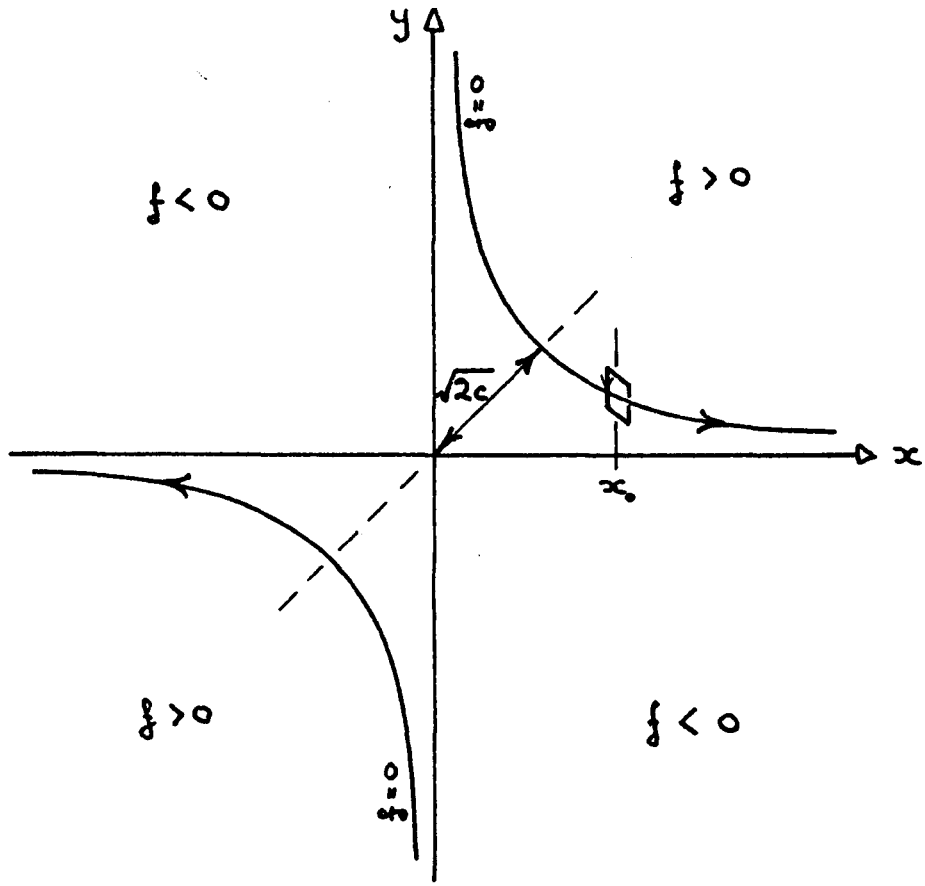


Figure 1.9 Hyperbolic Pair of Pure Edge Dislocations

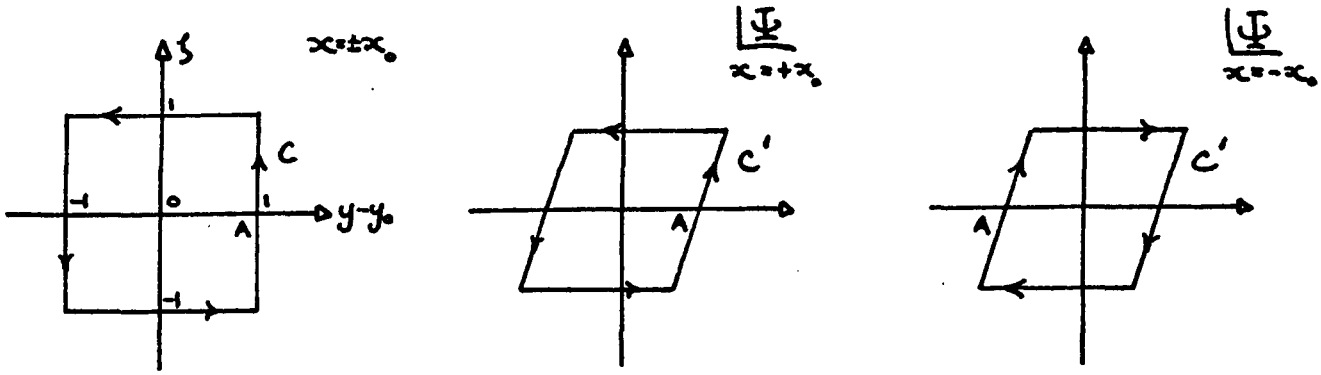


Figure 1.10 Circuits in Planes $x = +x_0$ and $x = -x_0$

Along the line $x = x_0$, if $f(x_0,y_0) = 0$ then $f(x_0,y) = x_0(y-y_0)$ which is obviously antisymmetric, and

$$\Psi(x_0,y,\zeta) = x_0(y-y_0) + \beta\zeta^{2n+1}$$

Let us take a circuit around the two branches of the hyperbola, at $x = +x_0$ and $x = -x_0$, as in fig. 1.10. At $+x_0$, C' encircles the origin once positively, and at $-x_0$ once negatively. We attach arrows to the dislocation line indicating the sense of the circuit C necessary to make C' positive. This model gives us two single dislocations of opposite sign, and by taking the limit $c \to 0$ we can make them coalesce at the origin, where we expect them to annihilate. It is interesting that the two dislocations cannot have the same sign, so that there is no chance of producing a double edge dislocation even at only a point. Let us take a circuit in the plane $x = y$ about the origin for $c = 0$. Then $f(x,y) = x^2$ which is clearly not antisymmetric, so that C' does not encircle the origin and the total dislocation strength enclosed is zero, as expected.

Let us examine the "local antisymmetry" condition further, by finding what behaviour of $f(x,y)$ about its zeros is allowed by the Laplace equation. For example, if $f(x,y)$ could locally have the form of a curved cylinder touching the xy plane, it would not be "locally antisymmetric" about its zero line. $f(x,y)$ must be the real part of an analytic function of $(x+iy)$ to satisfy the Laplace equation. Let us expand this about a point where $f(x,y) = 0$, taken as origin.

$$f(x,y) = \mathrm{Re}\ \{\ a(x+iy) + b(x+iy)^2 + c(x+iy)^3 + \ldots\ \}$$

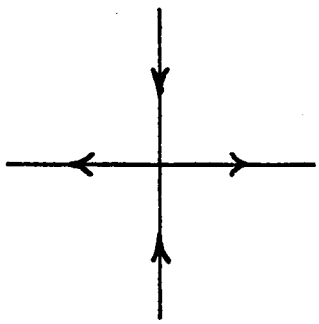
If $a \neq 0$, the local form is $(a_R x - a_I y)$ which is linear. $f(x,y)$ is locally antisymmetric and we have a simple isolated dislocation line.

If $a = 0$, $b \neq 0$, the local form is $b_R(x^2-y^2) - b_I 2xy$ which is a hyperbolic paraboloid (see HCV52, p15) with the saddle point at height zero. $b_R(x^2-y^2) - b_I 2xy = 0$ is the equation of two straight lines intersecting at the origin <u>at right angles</u>. We have already illustrated this case. At the origin $f(x,y)$ is locally symmetric, so that the total dislocation strength is zero, because two dislocations of opposite sign are touching at a point and cancelling.

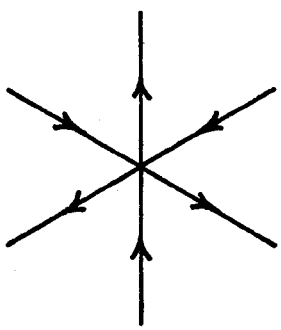
If $a = b = 0$, $c \neq 0$, the local form is $c_R(x^3-3xy^2) - c_I(3x^2y-y^3)$ which is a rotated form of $c_R(x^3-3xy^2)$. This is a monkey (or double)

saddle (see HCV52, p191) at height zero, and appears to produce three straight dislocation lines intersecting at the origin at 60^{o}. But it is really three dislocations lines with 60^{o} corners which meet at a point, as we see by adding a constant to f(x,y) to split them up. At the origin f(x,y) is locally antisymmetric so that the total dislocation strength is 1, because two of the dislocations cancel. Higher terms of f(x,y) similarly lead to more degenerate saddles.

We have established that a given number of pure edge dislocations can only touch each other in one configuration, as illustrated in fig. 1.11. Also critical points of f(x,y) can only be isolated (possibly degenerate) saddles, therefore the fact that f(x,y) satisfies the Laplace equation ensures that it is locally antisymmetric about its zeros, except where an even number of dislocations touch, when they must completely cancel.



Touching of 2 Static Pure Edge Dislocations



Touching of 3 Static Pure Edge Dislocations

Figure 1.11

1.5 Lorentz Transformation of Static Dislocations

One might imagine that if one viewed a dislocation from a frame of reference moving with respect to the frame in which the dislocation was generated, then it might appear to have a different character; for example, it might be possible to make a static dislocation move. If we try to do this by applying a Galilean transformation to a wavefunction satisfying the scalar wave equation in the rest frame, then it will not do so in the moving frame, since the scalar wave equation is not invariant under Galilean transformation. However, the scalar wave equation is invariant under Lorentz transformations. If we consider scalar waves whose wave velocity c is that of light, then a Lorentz transformation physically represents observing from a moving frame of reference. If the wave velocity is not that of light, then a Lorentz transformation is just a formal transformation which will generate a different solution of the original wave equation, which may show new behaviour.

First we show that a dislocation is a Lorentz invariant concept. Suppose L is a general Lorentz transformation from a frame S to a frame S' such that $x' = Lx$ for 4-vectors x in S and x' in S'. Since the (complex) wavefunction ψ is a Lorentz scalar, $\psi'(x') = \psi(x)$. The equation of a dislocation line in S is $\psi(x) = 0$ which implies $\psi'(x') = 0$. Therefore, if x lies on a dislocation line D in S then $x' = Lx$ lies on that dislocation line D' in S'. Take a circuit C around D in S, then the strength σ of the dislocation in S is given by

$$2\pi\sigma = \oint_C d\chi(x) = \oint_C \left(\frac{\partial\chi(x)}{\partial x}\right) dx \ .$$

But $$\frac{\partial\chi(x)}{\partial x} = \frac{\partial\chi'(x')}{\partial x'}\,\frac{dx'}{dx} = \frac{\partial\chi'(x')}{\partial x'}\,L$$

because phase χ is Lorentz scalar and $dx' = L\,dx$. Changing the integration variable to x' we have

$$2\pi\sigma = \oint_{C'} \left(\frac{\partial\chi'(x')}{\partial x'}\right) L\,dx = \oint_{C'} \left(\frac{\partial\chi'(x')}{\partial x'}\right) dx' = \oint_{C'} d\chi' = 2\pi\sigma'$$

Therefore dislocation strength is Lorentz invariant.

We shall use the notation $x^{\mu} \equiv (x,y,z,ct)$ where $\mu = 1,2,3,4$ with metric tensor $g^{\mu\nu} = \text{diag.}(1,1,1,-1)$ (e.g. see Rindler(1960), but note that his metric tensor is minus ours). S' will move with velocity v relative to S and $\gamma = (1 - v^2/c^2)^{-\frac{1}{2}}$. We shall always choose a coordinate system such that our basic unmodulated plane wave travels along the z axis. Suppose S' is moving in standard configuration (see Rindler, op.cit.) along the z axis of S, then the Lorentz transformation matrix is

$$\Lambda^{\mu}{}_{\nu}(v) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & -\gamma v/c \\ 0 & 0 & -\gamma v/c & \gamma \end{pmatrix}$$

The wave 4-vector $k^{\mu} = k(0,0,1,1)$ becomes $k'^{\mu} = \gamma(1-v/c)k(0,0,1,1)$ and is still along the z' axis, but the frequency has been Doppler shifted.

$$k^{\mu}x_{\mu} = k'^{\mu}x'_{\mu} \quad \text{(Lorentz scalar product)}$$

$$\Rightarrow kz-\omega t = k'z'-\omega' t'$$

$$\text{or } k\zeta = k'\zeta'$$

Consequently a static dislocation will always remain static under <u>any</u> Lorentz transformation. Also

$$x = x'$$

$$y = y'$$

$$z = \gamma(z'+vt')$$

$$ct = \gamma(z'v/c+ct')$$

from the inverse transformation. Then any wavefunction having the form

$$\psi(x,y,\zeta) = f(x,y,k\zeta) \tag{1.8}$$

in S has exactly the same form in S', viz.

$$\psi'(x',y',\zeta') = \psi(x,y,\zeta) = f(x',y',k'\zeta')$$

It seems reasonable to use k as a scale factor in the direction of travel of the plane waves, but not perpendicular to that direction, and therefore to choose the scale factors in the envelope functions such that we always write our canonical dislocations in the form (1.8). Then this transformation produces no effect other than a frequency shift.

If S' is moving in standard configuration along the x axis (say) of S, then

$$\Lambda^{\mu}{}_{\nu}(v) = \begin{pmatrix} \gamma & 0 & 0 & -\gamma v/c \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma v/c & 0 & 0 & \gamma \end{pmatrix}$$

and $k^{\mu} = k(0,0,1,1)$ transforms into $k'^{\mu} = k(-\gamma v/c, 0, 1, \gamma)$. The plane wave is now travelling in the x'z' plane at some angle θ to the z' axis, so we may write $k'^{\mu} = k'(-\sin\theta, 0, \cos\theta, 1)$ where $\cos\theta = 1/\gamma$, $\sin\theta = v/c$ and $k' = \gamma k$. The wave has suffered a frequency shift plus aberration. But we always use a coordinate system with z along $\underline{k}$, therefore we must rotate S' through an angle θ into S", such that

$$\begin{pmatrix} x'' \\ z'' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x' \\ z' \end{pmatrix}$$

as in fig. 1.12. Substituting the values of $\cos\theta$ and $\sin\theta$, the full transformation matrix from S' to S" is the spatial rotation

$$R^{\mu}{}_{\nu} = \begin{pmatrix} 1/\gamma & 0 & v/c & 0 \\ 0 & 1 & 0 & 0 \\ -v/c & 0 & 1/\gamma & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The full Lorentz transformation L from S to S" is $L^{\mu}{}_{\nu}(v) = R^{\mu}{}_{\sigma}\Lambda^{\sigma}{}_{\nu}(v)$. Then the inverse transformation is

$$(L^{-1})^{\mu}{}_{\nu} = \begin{pmatrix} 1 & 0 & -\gamma v/c & \gamma v/c \\ 0 & 1 & 0 & 0 \\ v/c & 0 & 1/\gamma & 0 \\ v/c & 0 & -\gamma v^2/c^2 & \gamma \end{pmatrix}$$

giving

$$x = x'' - \gamma v/c(z''-ct'')$$

$$y = y''$$

$$z = x''v/c + z''/\gamma$$

$$ct = x''v/c - z''\gamma v^2/c^2 + \gamma ct''$$

and once again

$$k(z-ct) = k''(z''-ct'')$$

or $k\zeta = k''\zeta''$ where $k'' = \gamma k$.

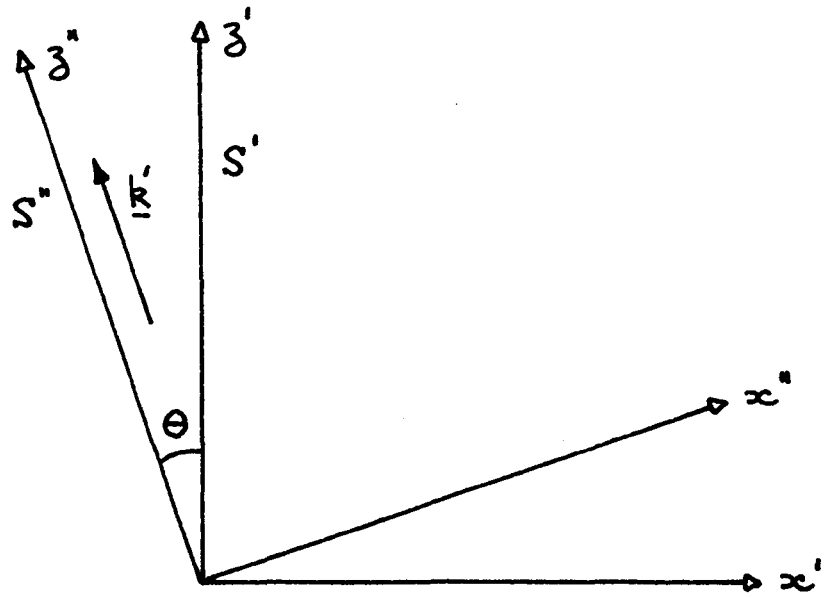


Figure 1.12 Transformation from S' to S"

So if

$$\psi(x,y,\zeta) = f(x,y,k\zeta)$$

then $\psi''(x'',y'',\zeta'') = f(x''-\zeta''\gamma v/c, y'', k''\zeta'')$.

The single strength mixed dislocation (1.4) may be written

$$\psi(x,y,\zeta) = (x + \beta_s y + \beta_e k\zeta)e^{ik\zeta}$$

which becomes

$$\psi''(x'',y'',\zeta'') = (x'' + \beta_s y'' + (\beta_e k''-\gamma v/c)\zeta'')e^{ik''\zeta''}$$

The equation of the dislocation line D becomes

$$x'' + \beta_{sR} y'' + (\beta_{eR} k''-\gamma v/c)\zeta'' = 0$$

$$\beta_{sI} y'' + \beta_{eI} k''\zeta'' = 0$$

i.e. the angle of the projection of D onto the $x''\zeta''$ plane has been changed by a Lorentz transformation along the x axis. The effect of this is to change the angle δ of D to the wavefronts (xy plane) and rotate it about $\underline{k}$, so generally the screw-edge character has been changed. If $\beta_e = 0$, we start with a pure screw and produce a mixed screw-edge in the $x''\zeta''$ plane, with $\tan\delta = c/\gamma v$. Conversely, if we choose our axes such that a mixed dislocation lies in the $x\zeta$ plane (i.e. $\beta_{eI} = 0$), and transform along the x axis, then if $\gamma v/c = \beta_{eR} k''$ we produce a pure screw. We require $v/c = \beta_{eR} k < 1$, $\Rightarrow \pi/4 < \delta \leqslant \pi/2$, which means that if $\beta_{eR} k$, and hence the degree of edge character of the original dislocation, is not too large, we can transform it into a pure screw. If β_s is real we start with a pure edge, and the dislocation line does not change at all. As $v \to c$, $\gamma \to \infty$ and any dislocation becomes more edge-like, but in the limit it would cease to be a dislocation.

The mixed multiple dislocation (1.5) may be written

$$\psi(x,y,\zeta) = \left[(x \pm iy) + \beta k\zeta\right]^s e^{ik\zeta}$$

which becomes

$$\psi''(x'',y'',\zeta'') = \left[(x'' \pm iy'') + (\beta k''-\gamma v/c)\zeta''\right]^s e^{ik''\zeta''}$$

and all the above comments apply, except that now the pure edge form never exists.

The general pure edge dislocation (1.7) may be written

$$\psi(x,y,\zeta) = \{f(x,y) + \beta(k\zeta)^{2n+1}\}e^{ik\zeta}$$

where $f(x,y)$ is a real solution of the Laplace equation, β is not pure real and n is a non-negative integer, which becomes

$$\psi''(x'',y'',\zeta'') = \{f(x''-\zeta''\gamma v/c,y'') + \beta(k''\zeta'')^{2n+1}\}e^{ik''\zeta''}$$

The equation of the dislocation line is

$$f(x''-\zeta''\gamma v/c,y'') = 0$$

$$(k''\zeta'')^{2n+1} = 0$$

i.e. $f(x'',y'') = 0$

$$\zeta'' = 0$$

hence it is generally true that pure edge dislocation lines do not change under this transformation.

A general homogeneous Lorentz transformation L may be decomposed into transformations Λ along the three axes of the original frame and a rotation R. If L is required to keep $\underline{k}$ along z, we can ensure this by arranging each Λ and R to keep $\underline{k}$ along z, i.e. R represents only a rotation in the xy plane at some stage. The two-dimensional Laplace equation is invariant under rotation, therefore all the canonical static dislocations are <u>form</u> invariant under R, and also as we have seen above, under each Λ. Hence the canonical static dislocations are <u>form</u> invariant under a general homogeneous Lorentz transformation, and remain static. Pure edges remain pure edges, but pure screws may become mixed, and some mixed dislocations may become pure screws.

1.6 Local Models for Moving Dislocations

The next three sections are based on section 7 of NB74. By moving we mean moving relative to the (unmodulated) wavefronts, and by analogy with crystal dislocations we call motion in the plane defined locally by the dislocation line and the wavenormal (the glide plane) glide, and motion perpendicular to this plane climb (see NB74 section 6). We again write our wavefunction as

$$\Psi(x,y,z,t)e^{ik\zeta},$$

but to obtain motion we require that z and t do not occur only in the combination $\zeta = z-ct$. We shall see that this necessitates quadratic terms in Ψ, so we shall take as our basic model the most general quadratic in x,y,z,t which satisfies the wave equation. Now

$$\left(\nabla^2 - \tfrac{1}{c^2}\tfrac{\partial^2}{\partial t^2}\right)\Psi(x,y,z,t)\,f(z-ct)$$

$$= f(z-ct)\left(\nabla^2 - \tfrac{1}{c^2}\tfrac{\partial^2}{\partial t^2}\right)\Psi(x,y,z,t) + 2f'(z-ct)\left(\tfrac{\partial}{\partial z} + \tfrac{1}{c}\tfrac{\partial}{\partial t}\right)\Psi(x,y,z,t)$$

where $f'(z) = \partial f/\partial z$, so taking $f(\zeta) = e^{ik\zeta}$ gives

$$\left(\nabla^2 - \tfrac{1}{c^2}\tfrac{\partial^2}{\partial t^2}\right)\Psi(x,y,z,t)\,e^{ik\zeta}$$

$$= e^{ik\zeta}\left[\left(\nabla^2 - \tfrac{1}{c^2}\tfrac{\partial^2}{\partial t^2}\right) + 2ik\left(\tfrac{\partial}{\partial z} + \tfrac{1}{c}\tfrac{\partial}{\partial t}\right)\right]\Psi(x,y,z,t)$$

We apply the above operator [⟷] to the general quadratic and equate coefficients to zero. Since this operator is invariant under rotations in the xy plane, we can always rotate the coordinates to remove the term xy, giving the general quadratic for Ψ as

$$\Psi(x,y,z,t) = \alpha x + \beta y + (\gamma x+\delta y)(z-ct) + \varepsilon(z-ikx^2) + \eta(ct-ikx^2)$$
$$+ \lambda(x^2-y^2) + \mu(z-ct)^2 + \nu \qquad (1.9)$$

where all coefficients may be complex. We can still add any function of $\zeta = z-ct$ to this: in particular we shall consider the term $\mu(z-ct)^2$ replaced by $\mu(z-ct)^3$.

This form for Ψ can obviously exhibit a wealth of behaviour, of which we can only hope to study some special cases. Wavefields are often "essentially two-dimensional" due to symmetry, and this is the case for the two cases we study in later chapters: one has cylindrical symmetry about the z axis, and the other has translation symmetry along the y axis. In such cases only pure edge dislocations may occur, so we shall consider models of these first.

1.7 Moving Pure Edge Dislocations

Pure edge dislocations must lie in planes perpendicular to the z axis, whose equations are $g(z,t) = 0$ where $g(z,t)$ is real, and this must be implied by $\Psi(x,y,z,t) = 0$. This can be arranged by making $\alpha,\beta,\gamma,\delta,\lambda$ pure real and $\eta_R = -\varepsilon_R$ giving

$$\text{real: } \alpha x + \beta y + (\gamma x+\delta y)\zeta + (\varepsilon_I+\eta_I)kx^2 + \lambda(x^2-y^2) + \varepsilon_R\zeta + \mu_R\zeta^2 + \nu_R = 0 \quad (1.10)$$

$$\text{imag: } \varepsilon_I z + \eta_I ct + \mu_I\zeta^2 + \nu_I = 0 \quad (1.11)$$

First, we consider the motion of the planes containing the dislocation lines, by solving (1.11) for $z(t)$, and also for $\zeta(t) = z(t)-ct$ which we need to substitute into (1.10). We can remove ν_I by shifting the zt origin.

$\mu_I = 0$ gives

$$z = -(\eta_I/\varepsilon_I)ct \qquad \Rightarrow \qquad \zeta = -(\eta_I/\varepsilon_I + 1)ct$$

Equation (1.11) is responsible for the glide component of the motion, and equation (1.10) for any climb component. We shall define the glide velocity to be $\partial\zeta/\partial t$ (not $\partial z/\partial t$ as used in NB74), and in the simple case above the glide velocity is constant and equal to

$$-(\eta_I/\varepsilon_I + 1),$$

which can take any value. If it is zero, then $\zeta = 0$ in equation (1.10), which becomes time independent. Therefore it is not possible within this model for pure edge dislocations to climb without gliding.

If $\mu_I \neq 0$ we may write (1.11) as

$$\varepsilon_I\zeta + (\varepsilon_I+\eta_I)ct + \mu_I\zeta^2 = 0$$

Then t as a function of ζ is just a parabola, which is easily plotted to show ζ as a function of t (fig.1.13). We can then add ct to ζ and plot z as a function of t (fig. 1.14). With the coefficients as shown we start out with two planes of edge dislocations moving with the wave. As time increases, they approach each other with increasing speed. The front one stops and then moves backwards, until the two planes meet. At this point the dislocations in the two planes lie along the same curves, and since they have opposite sign they annihilate. μ_I negative would produce a similar creation event.

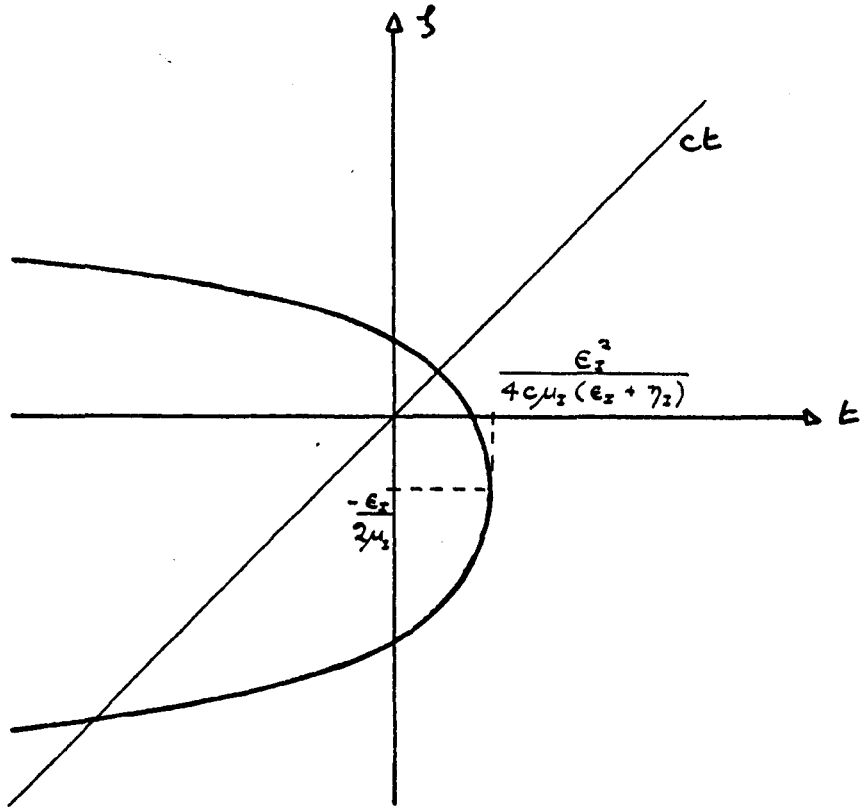


Figure 1.13 Quadratic Glide relative to Wavefronts

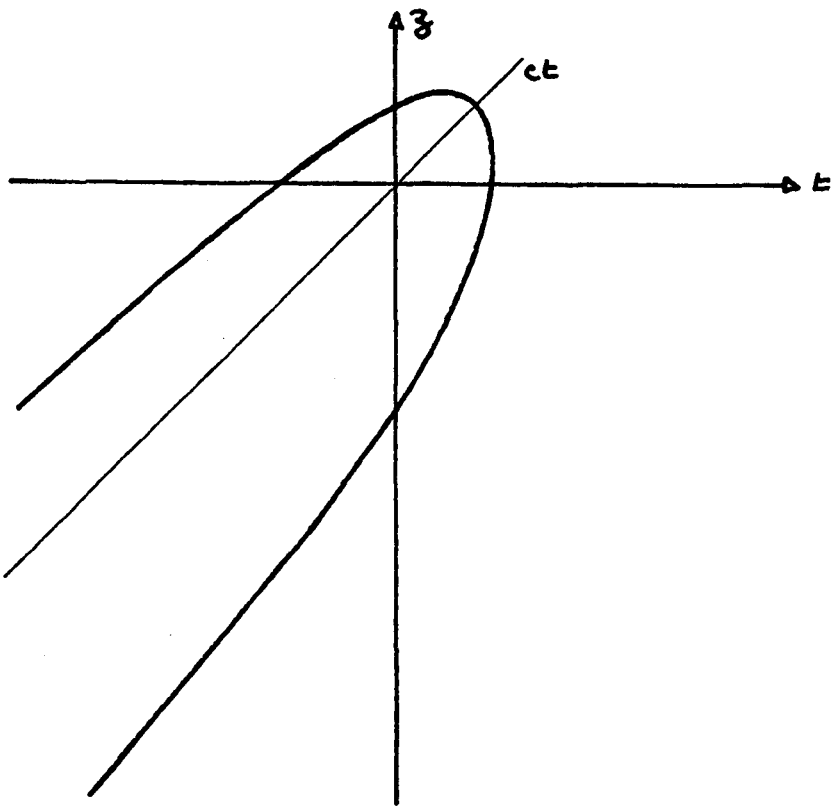


Figure 1.14 Quadratic Glide relative to Source

"Quadratic glide" produces birth/death events. Next we shall consider "cubic glide", where we replace the imaginary part of $\mu(z-ct)^2$ by $\mu_I(z-ct)^3$ and rewrite equation (1.11) as

$$\varepsilon_I\zeta + (\varepsilon_I + \eta_I)ct + \mu_I\zeta^3 = 0 \qquad (1.12)$$

Again this is most easily plotted by regarding t as a function of ζ, and then adding ct to ζ to plot z (figs. 1.15 & 1.16). These figures assume that

$$\varepsilon_I + \eta_I > 0 \quad \text{and} \quad |\varepsilon_I/(\varepsilon_I+\eta_I)| < 1.$$

In the distant past and distant future we have one dislocation, whose main behaviour is to glide backwards if $\mu_I > 0$ or forwards if $\mu_I < 0$. If $\varepsilon_I\mu_I > 0$ there is only one dislocation, and this is the whole story. But if $\varepsilon_I\mu_I < 0$ there is a transition regime with three dislocations. A pair of dislocations is born behind our original dislocation if $\mu_I > 0$, or in front if $\mu_I < 0$. The member of this pair with opposite sign approaches the original dislocation and annihilates it, leaving the other member of the pair to carry on instead of the original dislocation, as though the original dislocation had "skipped" over the intermediate dislocation. We shall call this process "dislocation skip": it provides a mechanism for very rapid gliding, and we shall see it in action in a later chapter. $\varepsilon_I = 0$ is the skip threshold: for $\varepsilon_I\mu_I$ just greater than 0 there is only rapid glide, for $\varepsilon_I\mu_I$ just less than 0 there is rapid glide plus skip. Figure 1.16 just represents zt sections through kinked 4-trajectories, and whether or not skip occurs may be changed by a small local rotation in space-time of this surface; it is due to one trajectory being interpreted in different ways because the time coordinate is singled out to have a different significance from the space coordinates. We could attach a direction to the graphs in figure 1.16, as shown by the arrows, and a tangent vector $\underline{s}$ pointing in the direction of the graph. If $\underline{t}$ is a vector along the positive t axis, then where $\underline{s}.\underline{t} > 0$ we have a positive dislocation, where $\underline{s}.\underline{t} < 0$ we have a negative dislocation, and where $\underline{s}.\underline{t} = 0$ we have no dislocation (i.e. a birth/death event). This classification is obviously not invariant under general space-time rotations (i.e. Lorentz transformations).

This description of the behaviour of dislocations is reminiscent of the Feynman diagram description of the interaction of elementary

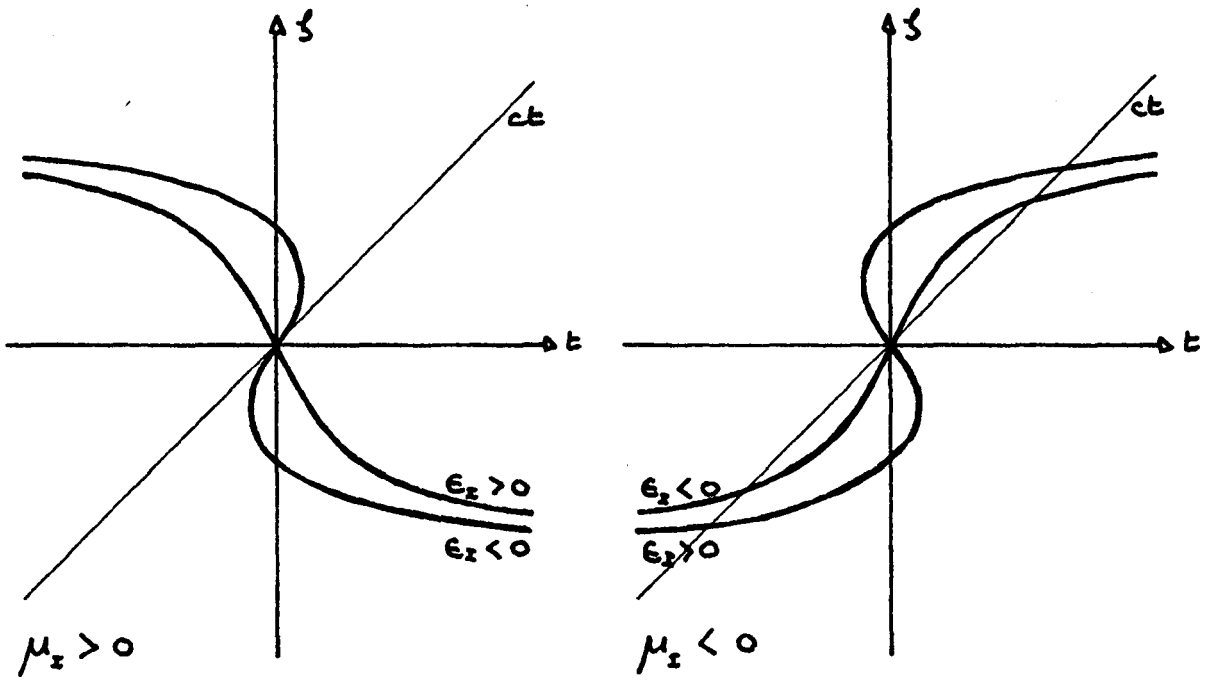


Figure 1.15 Cubic Glide relative to Wavefronts

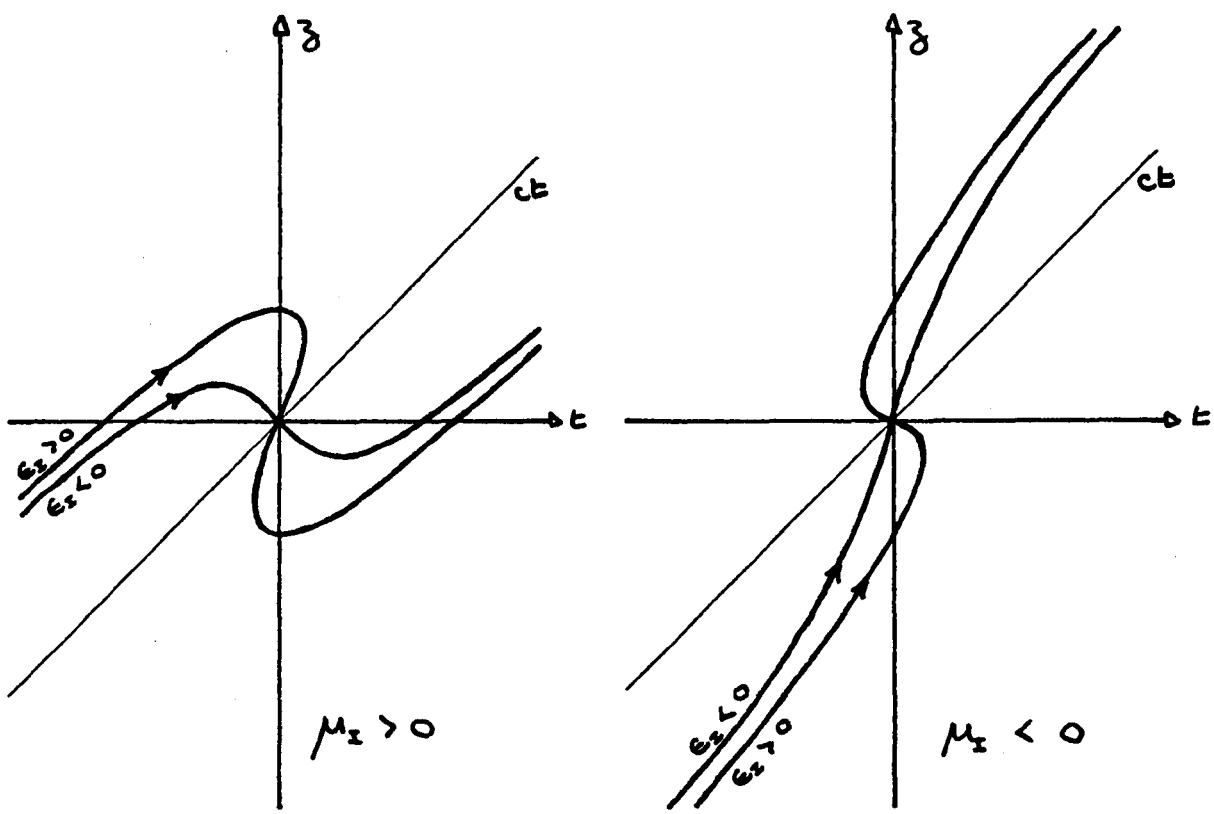


Figure 1.16 Cubic Glide relative to Source

particles. In particular, skip is analogous to the mechanism underlying Zitterbewegung (e.g. see Bjorken and Drell (1964)). Particle-antiparticle pairs are continually being created out of the vacuum, and then annihilating back into it. A particle travelling through the vacuum may annihilate one of these antiparticles, leaving the particle created from the vacuum to carry on instead of the original particle. This particle may then do the same again. The net effect is that a particle appears to "jiggle about" with instantaneous velocity c, although its average velocity is the correct "classical" velocity. This behaviour is predicted by the Dirac equation, and what we have described above is the field theoretical explanation of it.

If equation (1.11) was an nth order polynomial the trajectory could have up to (n-1) kinks and the dislocation could skip up to Int{(n-1)/2} times. In fact, there is no limit to the number of times a dislocation may skip, and it really can exhibit behaviour similar to Zitterbewegung if the wavefunction is sufficiently complicated. As a simple example replace ζ^2 in (1.11) by sin ζ giving

$$\varepsilon_I\zeta + (\varepsilon_I+\eta_I)ct + \mu_I \sin\zeta = 0$$

which is plotted in figure 1.17 for positive coefficients.

We have studied the behaviour of the planes containing the pure edge dislocations, and now we turn to the behaviour of the dislocations within one of these planes. We study the solutions of equation (1.10) as a function of ζ, and then to construct a complete set of dislocations we substitute values of $\zeta(t)$ as derived previously. The simplest case is pure glide when equation (1.10) is

$$\alpha x + \beta y + \{(\varepsilon_I+\eta_I)k + \lambda\}x^2 - \lambda y^2 + \nu_R = 0$$

If $(\varepsilon_I+\eta_I) = 0$ this is a hyperbola, and equation (1.11) (and its variations) gives ζ = constant(s), so we are back to static dislocations. Otherwise it can be any conic section. $\lambda = 0$ gives a parabola, and if also $\beta = 0$ this degenerates into two parallel straight lines as in equations (42) and (44) of NB74. If $\lambda\{(\varepsilon_I+\eta_I)k + \lambda\} < 0$ we have an ellipse, and in the special case $\lambda = -(\varepsilon_I+\eta_I)k/2$ this is the circular edge dislocation loop of equation (47) of NB74.

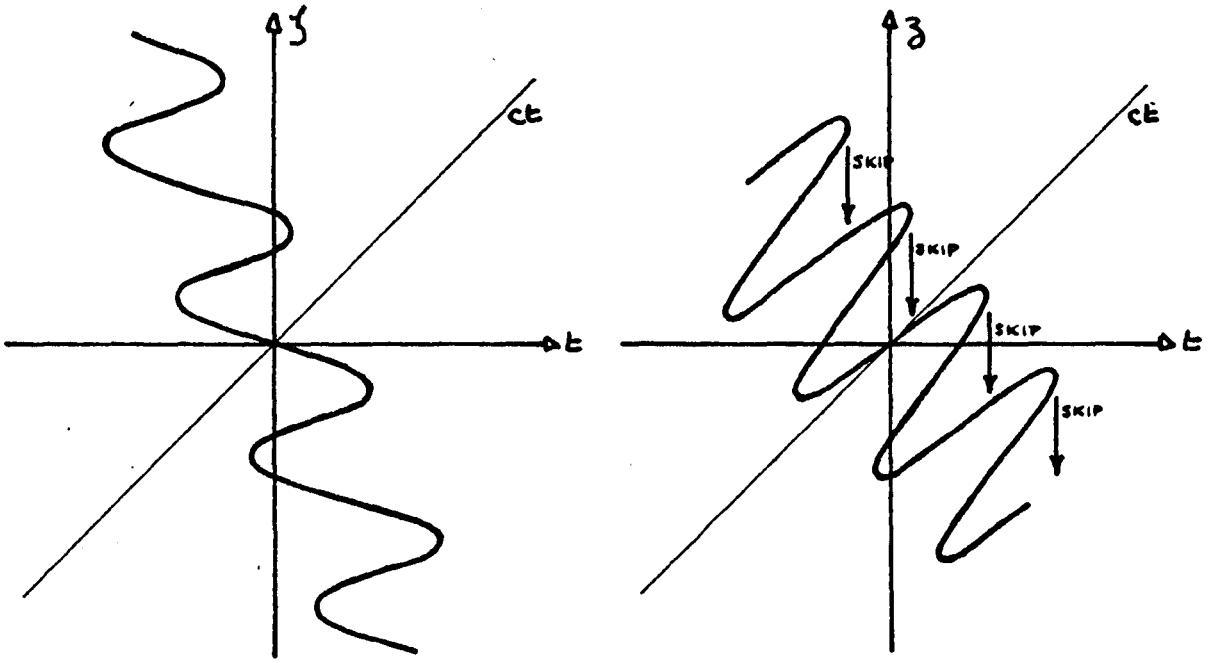
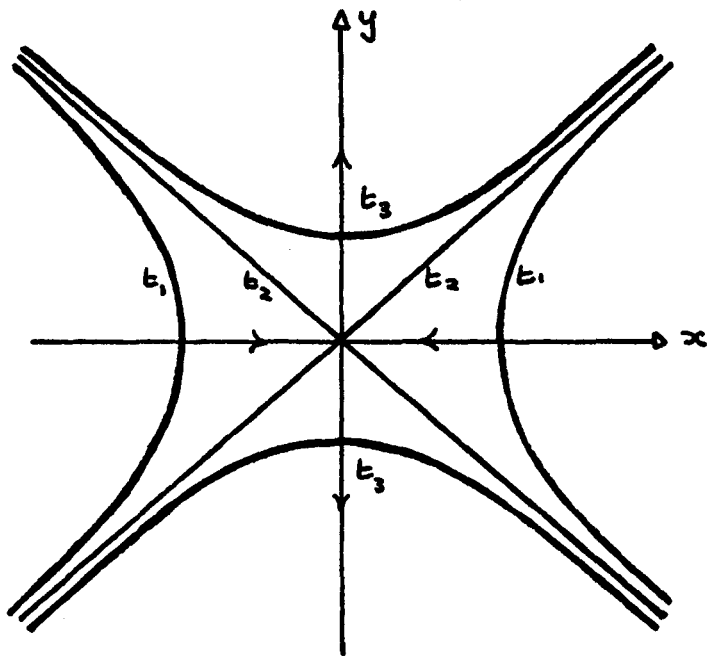


Figure 1.17 "Zitterbewegung" of Dislocations



Times $t_1 < t_2 < t_3$

Figure 1.18 Hyperbolic Collision of Edge Dislocations

The more general form of equation (1.10) allows climb, and we may write it as

$$\{(\varepsilon_I+\eta_I)k + \lambda\}\{x - X(\zeta)\}^2 - \lambda\{y - Y(\zeta)\}^2 + \rho(\zeta) = 0$$

where $X(\zeta)$ and $Y(\zeta)$ are linear, although $X(\zeta)$, $Y(\zeta)$ and $\rho(\zeta)$ may be arbitrary functions in general and the wave equation is still satisfied. The effect of $X(\zeta)$ and $Y(\zeta)$ is to translate the conic with time. More interesting is the effect of $\rho(\zeta)$ which causes the "radius" of the conic to vary. An ellipse may grow from a point or shrink to a point, representing a puncture in a wavefront appearing and opening, or closing and disappearing. We shall see an example of a circular puncture appearing and opening in a later chapter. Or the two branches of a hyperbola may coalesce and pull apart at right angles, as in fig. 1.18. One would expect this behaviour from the discussion in section 4. It should also be compared with equation (49) of NB74 which shows a different type of collision of two edge dislocations.

We have not said much about the actual trajectories of these dislocations, but before leaving this section we shall construct a dislocation with a particular stepped trajectory, because we shall come across it in our analysis of an experiment in a later chapter. We arrange equation (1.10) to have the form

$$(\varepsilon_I+\eta_I)kx^2 - \mu_I(\zeta+p)^2 = 0$$

giving straight pure edge dislocations parallel to the y axis, satisfying

$$x = \pm\sqrt{\frac{\mu_I}{(\varepsilon_I+\eta_I)k}}\,(\zeta+p) \qquad (1.13)$$

Instead of equation (1.11) we take the cubic glide form (1.12) with $\mu_I > 0$ and $\eta_I = 0$ giving

$$t = \frac{-1}{c}\left(\frac{\mu_I}{\varepsilon_I}\,\zeta^3 + \zeta\right).$$

$\zeta(t)$ and $z(t)$ are plotted in fig. 1.19. From these two graphs and equation (1.13) we can plot the projection on the xz plane, fig. 1.20, which is a symmetrical pair of stepped dislocation trajectories.

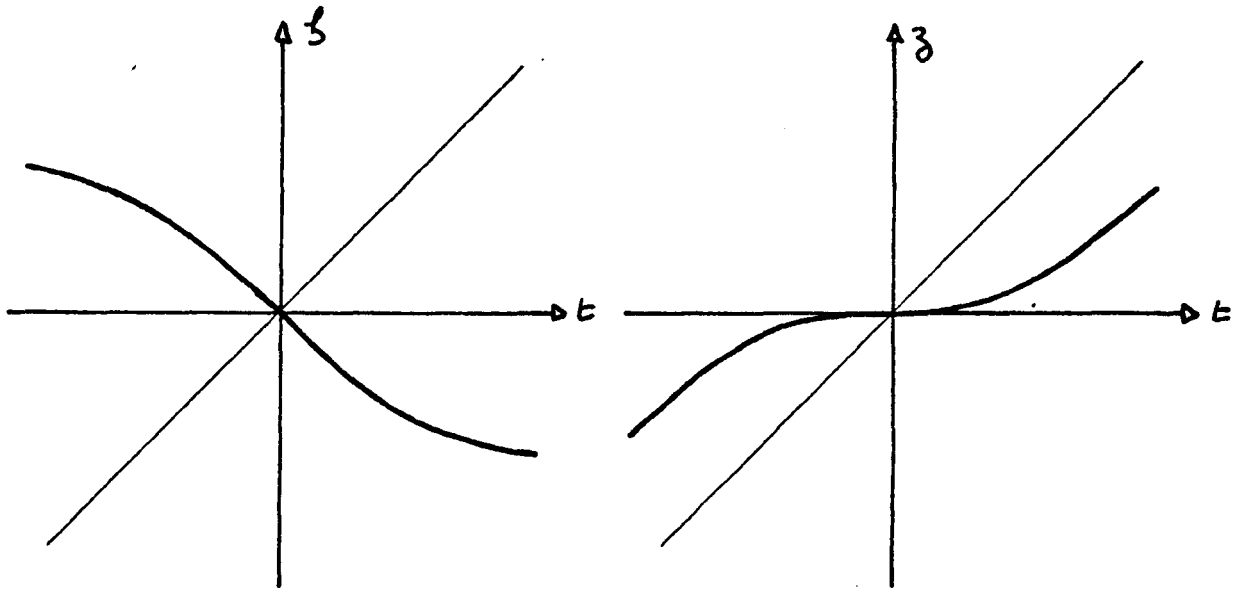


Figure 1.19 Cubic Glide with dz/dt = 0 at Origin

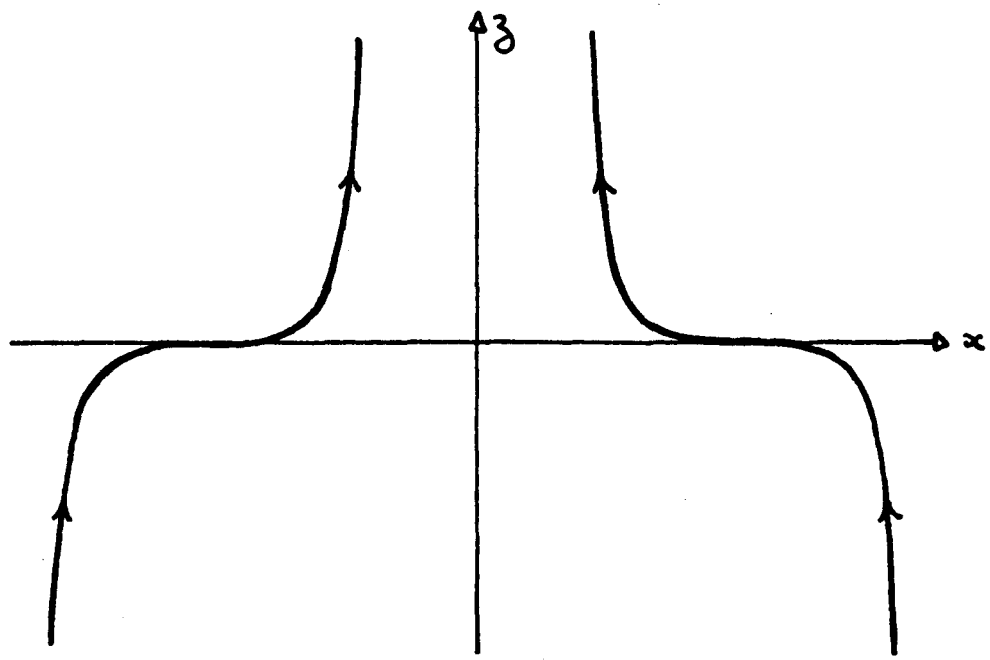


Figure 1.20 Pair of Stepped Dislocation Trajectories

1.8 Moving Screw and Mixed Dislocations

A pure screw dislocation must be parallel to the z axis, therefore Ψ must have the form

$$\Psi(x,y,z,t) = \alpha x + \beta y + \eta(ct-ikx^2) + \lambda(x^2-y^2) + \nu.$$

The equation of the dislocation line is given by

$$\alpha_R x + \beta_R y + \eta_R ct + (\eta_I k+\lambda_R)x^2 - \lambda_R y^2 + \nu_R = 0 \qquad (1.14)$$

$$\alpha_I x + \beta_I y + \eta_I ct + (-\eta_R k+\lambda_I)x^2 - \lambda_I y^2 + \nu_I = 0 \qquad (1.15)$$

Each of these equations may be a straight line or a (possibly degenerate) conic varying with time. They can only both be straight lines if they are both independent of t, and we are back to the single static dislocation. Let us make (1.14) a straight line by setting $\eta_I = \lambda_R = 0$. Then (1.14) and (1.15) become

$$\alpha_R x + \beta_R y + \eta_R ct + \nu_R = 0 \qquad (1.16)$$

$$\alpha_I x + \beta_I y + (\lambda_I-\eta_R k)x^2 - \lambda_I y^2 + \nu_I = 0 \qquad (1.17)$$

These equations give a straight line which sweeps across an arbitrary fixed conic as time varies, and the dislocation lines lie at the intersection of the two curves. This gives us generally 0 or 2 dislocations. If (1.17) is a parabola we have equation (55) of NB74, and if the parabola degenerates into two straight lines we have equation (54) of NB74. If (1.16) is parallel to the axis of the parabola, we only have one dislocation gliding along the parabola as in equation (53) of NB74. This is the only case which gives a single moving screw dislocation in this model, and is a very special case. The behaviour of the pair of dislocations produced by the straight line sweeping over an ellipse or hyperbola is obvious.

If we eliminate t between (1.14) and (1.15) we always get the equation of a conic, hence the dislocations always move round a conic. But in the most general case they lie on the intersections of two time dependent conics, so there may be 0,2 or 4 dislocations, and the possible behaviour is fairly obvious. As a simple example we will consider two parabolae at right angles to each other. We take $\lambda_R = 0$ in (1.14) and $\lambda_I = \eta_R k$ in (1.15) giving

$$\alpha_R x + \beta_R y + \eta_R ct + \eta_I kx^2 + \nu_R = 0 \tag{1.18}$$

$$\alpha_I x + \beta_I y + \eta_I ct - \eta_R ky^2 + \nu_I = 0 \tag{1.19}$$

These are plotted in fig. 1.21 for a suitable choice of coefficients, and the direction of motion in time shown by the arrows. As plotted these graphs give 4 dislocations. These will separate along y slowly, but the top pair will coalesce along x followed by the bottom pair. Between them the 4 dislocations will trace out the elliptical trajectory shown dashed.

Pure screw dislocations are probably the simplest because they must be straight lines, whereas pure edge dislocations are plane curves, and mixed dislocations may be twisted space curves. For completeness, we conclude this section with a very simple example of a moving mixed dislocation. We take

$$\Psi(x,y,z,t) = i\alpha x + \beta y + i\varepsilon(z-ikx^2) + i\eta(ct-ikx^2)$$

with $\alpha,\beta,\varepsilon,\eta$ real. The equation of the dislocation line is

$$\text{real: } \beta y + (\varepsilon+\eta)kx^2 = 0 \tag{1.20}$$

$$\text{imag: } \alpha x + \varepsilon z + \eta ct = 0 \tag{1.21}$$

Equation (1.20) is a parabolic cylinder parallel to the z axis, and (1.21) is a plane parallel to the y axis. It intersects the parabolic cylinder in a parabola which moves along the z axis as time varies. This parabola is the dislocation line which is of mixed screw-edge type along the whole of its length. A general mixed dislocation line could have pure edge or screw points along it, and could change its shape as it moves with time.

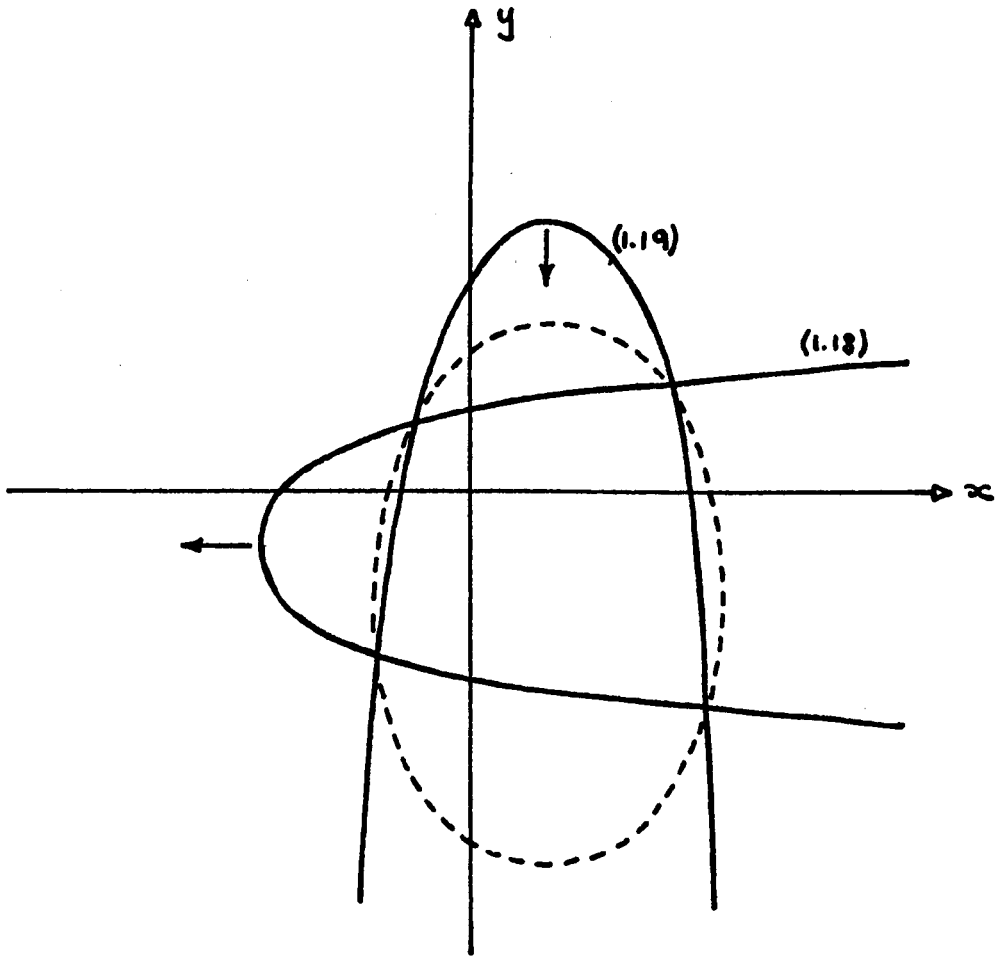


Figure 1.21 Possible Motion of Pure Screw Dislocations around an Ellipse (dislocations are at intersections of parabolae)

1.9 Dislocations in Electromagnetic Waves

Although we consider mainly scalar waves in this work, as an indication of the generality of the concept of wavefront dislocations, we shall construct a simple dislocated electromagnetic wave to prove that they can exist. Consequently we have to work with a pair of complex 3-vector wavefunctions $\underline{E}$ and $\underline{B}$ (or we could work with the field 4-tensor). We could construct a vector potential and deduce $\underline{E}$ and $\underline{B}$ from it, but this is indirect if one is specifically trying to construct observable electromagnetic fields with certain properties.

We shall work in free space, where $\underline{E}$ and $\underline{B}$ must satisfy the Maxwell equations:

$$\nabla.\underline{E} = 0$$

$$\nabla.\underline{B} = 0$$

$$\nabla\wedge\underline{E} = - \partial\underline{B}/\partial t$$

$$\nabla\wedge\underline{B} = 1/c^2 \, \partial\underline{E}/\partial t$$

which imply that $\underline{E}$ and $\underline{B}$ satisfy the wave equation (e.g. see Bleaney & Bleaney (1965)). We shall choose a field $\underline{E}$ which satisfies the wave equation and constrain it to satisfy $\nabla.\underline{E} = 0$. From it we shall deduce $\underline{B}$ using $\underline{B} = - \int dt \, \nabla\wedge\underline{E}$, so that $\underline{B}$ automatically satisfies $\nabla.\underline{B} = 0$ and the wave equation. We shall then further constrain the fields so that $\underline{B}$ has the required form for the dislocation also.

When discussing dislocations we are concerned with the phase of a wavefunction, and for an arbitrary complex scalar wavefunction, this is well defined. But if a vector wavefunction has arbitrary complex components, the phase of the wavefunction is not well defined. It can be made meaningful if the variable part of the phase of each component, at least, is the same, as in the complex representation of an elliptically polarised plane wave. This implies that every component of the wavefunction has the same dislocation. For a dislocated electromagnetic wave we shall require that every non-zero component of $\underline{E}$ and $\underline{B}$ has the same dislocation. It is possible for different components to have different dislocations, or for some components to have no dislocations, but we shall ignore such possibilities as being not true vector-wave dislocations. We require that the electromagnetic field be zero along a dislocation line.

We shall only consider straight static dislocations, and we shall take as our basic local model

$$\underline{E}(x,y,z,t) = \underline{E}_o \Psi(x,y,\zeta) e^{ik\zeta}$$

$$\underline{B}(x,y,z,t) = \underline{B}_o \Psi(x,y,\zeta) e^{ik\zeta}$$

where $\underline{E}_o$ and $\underline{B}_o$ are complex constant vectors, and $\zeta = z-ct$. Based on our previous work on scalar dislocations we take

$$\Psi(x,y,\zeta) = \left[x + \beta_s y + \beta_e \zeta\right]^s$$

where β_s and β_e are not both real, and $\beta_s = \pm i$ unless $s = 1$, from section 3. If we can satisfy Maxwell's equations, we will have a straight static dislocation of strength s. $\underline{E}$ and $\underline{B}$ have been constructed to satisfy the wave equation, but

$$\nabla.\underline{E} = \{E_x\, \partial/\partial x + E_y\, \partial/\partial y + E_z\, \partial/\partial z\}\Psi e^{ik\zeta}$$

$$= \{E_x s + E_y s\beta_s + E_z(s\beta_e + ik[x + \beta_s y + \beta_e \zeta])\}[\;\cdot\;]^{s-1} e^{ik\zeta}$$

if $\underline{E}_o = (E_x, E_y, E_z)$. $\nabla.\underline{E} = 0$ for all x,y,ζ requires $E_z = 0$ and $E_x + E_y\beta_s = 0$. Therefore

$$\underline{E}_o = (1, -1/\beta_s, 0) \quad \text{(taking } E_x = 1\text{)}$$

and $$\nabla\wedge\underline{E} = \left(\frac{1}{\beta_s}\frac{\partial}{\partial z}\,,\ \frac{\partial}{\partial z}\,,\ -\left\{\frac{1}{\beta_s}\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right\}\right)\Psi e^{ik\zeta}$$

Then

$$\underline{B} = \frac{1}{c}\int d\zeta\ \nabla\wedge\underline{E}$$

$$= \frac{1}{c}\left(\frac{1}{\beta_s}\Psi e^{ik\zeta}\,,\ \Psi e^{ik\zeta}\,,\ -\frac{1}{\beta_e}\left\{\frac{1}{\beta_s} + \beta_s\right\}\int d\zeta\ \frac{\partial\Psi}{\partial\zeta}\, e^{ik\zeta}\right)$$

where we have set the constants of integration equal to zero, because we want to factor out $\Psi e^{ik\zeta}$. But we can only do this if we make $B_z = 0$ because otherwise it has the wrong form. We therefore require

$$\left\{\frac{1}{\beta_s} + \beta_s\right\} = 0 \quad \Rightarrow \quad \beta_s = \pm i \quad \text{for all } s.$$

We have constructed the electromagnetic wave

$$\underline{E} = (1,\pm i,0)\left[x \pm iy + \beta\zeta\right]^s e^{ik\zeta}$$

$$\underline{B} = \frac{\mp i}{c}(1,\pm i,0)\left[x \pm iy + \beta\zeta\right]^s e^{ik\zeta} = \frac{\mp i}{c}\,\underline{E}$$

This is a circularly polarised plane wave along the z axis, modulated to produce a mixed screw-edge dislocation of strength s, as discussed earlier for scalar waves.

Let us investigate the physical consequences of this dislocation, by finding its effect on the energy of the wave. The energy density U is given by

$$U = \varepsilon_o \tfrac{1}{2}(\underline{E}^2 + c^2\underline{B}^2)$$

and the energy flux is given by the Poynting vector

$$\underline{P} = \varepsilon_o c^2 \underline{E}\wedge\underline{B},$$

remembering that we must use real fields. Let us write

$$\psi_\pm(x,y,\zeta) = \left[x \pm iy + \beta\zeta\right]^s e^{ik\zeta}$$

Then the real fields are

$$\underline{E}_R = (\mathrm{Re}\psi_\pm\ ,\ \mp\mathrm{Im}\psi_\pm\ ,\ 0)$$

$$c\underline{B}_R = (\pm\mathrm{Im}\psi_\pm\ ,\ \mathrm{Re}\psi_\pm\ ,\ 0)$$

and $\underline{E}_R.\underline{E}_R = c^2\underline{B}_R.\underline{B}_R = |\psi_\pm|^2$

$$\underline{E}_R\wedge c\underline{B}_R = (0\ ,\ 0\ ,\ |\psi_\pm|^2)$$

so that $U(x,y,z,t) = \varepsilon_o|\psi_\pm|^2$ and $\underline{P}(x,y,z,t) = (0,0,cU(x,y,z,t))$.

The energy flux is just the energy density transported in the direction of propagation of the wave, just as for an unmodulated plane wave and contrary, perhaps, to one's hopes there is no circulation of energy (and hence momentum) around the dislocation line. The energy density increases from zero as one moves away from the dislocation line, and the contour surfaces of U are elliptic cylinders centred on the dislocation line, such that their xy sections are circular. At fixed time, in a plane z = constant, $U = \varepsilon_o r^{2s}$ where r is the distance from where the dislocation line passes through the plane.

These calculations suggest that the physical effects of wavefront dislocations in terms of energy, momentum etc. are not very significant. Indeed, there are situations in other models, such as sound waves, where the energy density does not even go to zero on the dislocation line. On the other hand the consequences of dislocations in quantum mechanical wavefunctions could be very significant (see the comments at the end of the prologue), being related to quantised magnetic flux lines, for example. Considerably more research into the physics of dislocations, as opposed to the mathematics, is needed.

CHAPTER 2

THE ACOUSTIC RADIATOR

2.1 Introduction

The aim of this chapter is to study (theoretically) dislocations arising in a fairly realistic physical wavefield. Our wavefield will be the soundfield generated by a rigid plane circular piston vibrating in an infinite planar baffle, which we shall drive with quasimonochromatic pulses having a Gaussian envelope. This model is the one usually taken to represent circular acoustic transducers, despite the fact that in practice they are rarely mounted in infinite planar baffles. Nevertheless it is considered to be among the simplest models for which an exact formula for the soundfield can be found, and which is physically realizable to a good approximation. The formula takes the form of a one-dimensional integral for which the analytical solution is not known, and consequently the wavefunction has to be evaluated numerically in general. We begin by deriving an exact expression for the wavefunction when the radiator is driven with an arbitrary time dependence. The method and result are believed to be slightly different from any published.

We consider sound waves in a semi-infinite perfect acoustic fluid such that the fluid velocity $\underline{u}$ is irrotational. Then we may take as our wavefunction the scalar velocity potential ψ such that $\underline{u} = -\nabla\psi$. Standard hydrodynamics shows that ψ satisfies the wave equation

$$\nabla^2\psi - 1/c^2 \, \partial^2\psi/\partial t^2 = 0$$

where c (the sound velocity) is a constant for the fluid, and the acoustic pressure fluctuation δp and density fluctuation δd due to the sound wave are given by

$$\delta p = d \, \partial\psi/\partial t \qquad \text{and} \qquad \delta d = \delta p/c^2$$

where d is the mean density, if the wave amplitude is small. (e.g.see Coulson(1965), Rutherford(1965), Morse(1948), Rschevkin(1963), Rayleigh (1878)). Hence a knowledge of ψ completely specifies the sound wave.

The fluid in the region $z > 0$ is bounded by the plane $z = 0$ apart from a disc of radius a centred on the origin, which is at a height $F(t)$ (fig. 2.1). Thus if $\underline{R}_0$ is a (two-dimensional) vector

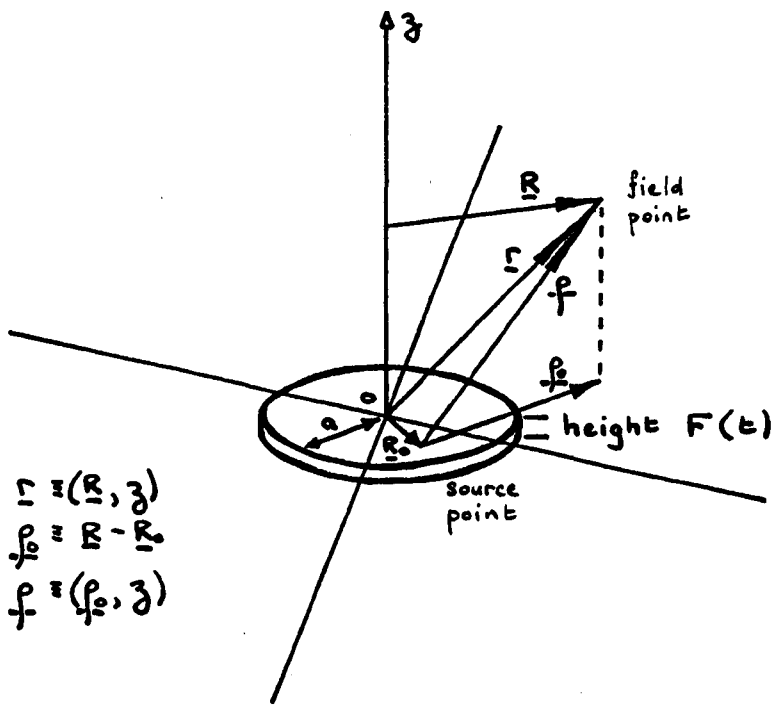


Figure 2.1 Coordinates for the Circular Piston Radiator

in the baffle plane, then the equation of the boundary surface is

$$z = G(\underline{R}_0)F(t) \tag{2.1}$$

where $G(\underline{R}_0) = \theta(a - |\underline{R}_0|)$ ($\theta(x)$ is the Heaviside unit step function). The boundary condition on the fluid is that the normal velocity of the fluid equals the normal velocity of the boundary surface. Then assuming the displacement of the boundary surface to be small,

$$\underline{u}_z\big|_{z=0} = -\partial\psi/\partial z\big|_{z=0} \simeq G(\underline{R}_0)F'(t)$$

which is known over the whole surface. It uniquely specifies the wavefunction in the whole half-space via Rayleigh's formula (for a plane boundary only. See Rayleigh (1878), p95, eqn.(3) for the CW case and p96, eqn.(8) for the general pulse case, but note that he defines the velocity potential with opposite sign. For the problem of the general boundary surface, see p95, eqn.(1), and also Farn & Huang (1968)):

$$\psi(\underline{r},t) = -\frac{1}{2\pi}\int_{\text{plane boundary}} \frac{d^2\underline{R}_0}{\rho}\,\frac{\partial\psi}{\partial z}(t-\rho/c) = \frac{1}{2\pi}\int_{\text{baffle plane}} d^2\underline{R}_0\,\frac{G(\underline{R}_0)\,F'(t-\rho/c)}{\rho}$$

where $\rho = (|\underline{R} - \underline{R}_0|^2 + z^2)^{\frac{1}{2}}$ is the distance from the source point at $\underline{R}_0$ to the field point at $\underline{r} = (\underline{R},z)$. Another derivation of this formula is given in appendix A2.1.

This is a Green function solution of the inhomogeneous wave equation in the absence of boundaries (the driving term now representing the baffle plane), with the integrand representing the "Huygens wavelet" from an element of the radiator (see Jackson (1962), pp183-188). However, it is worth noting that this problem is essentially different from that of diffraction by a circular aperture (contrary to the suggestion of some authors, e.g. Huntington et al.(1948) and Marini & Rivenez (1974)), where the wavefunction is not known over the whole boundary surface, and it is usually necessary to make the Kirchhoff approximation (see Jackson (1962), pp280-283). Use of Kirchhoff's surface-integral representation then leads to an approximate solution similar to that above, which is actually not self-consistent (see Levine & Schwinger (1948) and Spence (1949)), although Rayleigh's formula may be derived exactly using Kirchhoff's representation plus some additional physics (see Rschevkin (1963), chap.XI).

In order to evaluate the integral we change the variable of integration to ρ, the distance of the field point from the emitting point, by using the Dirac δ-function:

$$\psi(\underline{r},t) = \frac{1}{2\pi}\int_S d^2\underline{R}_0 \frac{G(\underline{R}_0)}{\rho(\underline{R}_0,\underline{r})} F'(t-\rho(\underline{R}_0,\underline{r})/c)$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\rho' \int_S d^2\underline{R}_0 \frac{G(\underline{R}_0)}{\rho(\underline{R}_0,\underline{r})} F'(t-\rho'/c)\,\delta(\rho'-\rho(\underline{R}_0,\underline{r}))$$

$$= \int_{-\infty}^{\infty} d\rho' F'(t-\rho'/c)\, H(\underline{r},\rho') \qquad (2.2)$$

where $$H(\underline{r},\rho') = \frac{1}{2\pi}\int_S d^2\underline{R}_0 \frac{G(\underline{R}_0)}{\rho(\underline{R}_0,\underline{r})}\,\delta(\rho'-\rho(\underline{R}_0,\underline{r})) \qquad (2.3)$$

is the "geometry function" for the radiator, so called because it is independent of the movement of the radiator in time, if this can be decoupled from the spatial variation as in eqn.(2.1). The wavefunction is the convolution of the velocity pulse function F'(t) with the geometry function $H(\underline{r},\rho)$. Note that if $F'(t) = \delta(t)$ then $\psi(\underline{r},t) = H(\underline{r},ct)$, hence $H(\underline{r},\rho)$ is the impulse response function as used by previous authors (e.g. Farn & Huang(1968), Oberhettinger(1961), Stepanishen(1971 a,b,c), Lockwood & Willette(1973)). Equation (2.3) shows that $H(\underline{r},\rho')$ only contributes to $\psi(\underline{r},t)$ when ρ' takes a value which is a physical distance from $\underline{r}$ to the radiating surface (otherwise the δ-function is zero). If we can find an analytical expression for $H(\underline{r},\rho)$ we only need do the one-dimensional integral (2.2) for the pulse functions of interest. Otherwise $H(\underline{r},\rho)$ could be tabulated once for a given radiator configuration, and used to compute the wavefunction for different pulse functions.

We shall simplify the geometry function as much as possible for a general circularly symmetric radiator excitation, and then specialise to the rigid plane circular piston vibrating in a fixed infinite planar baffle.

2.2 Evaluation of the Geometry Function for a General Circularly Symmetric Radiator Excitation

If $G(\underline{R}_o) = G(R_o)$ we should be able to do the angular integration. We can now replace $\underline{r}$ by the two cylindrical polar coordinates (R,z). If $\underline{R}.\underline{R}_o = RR_o\cos\theta_o$ then

$$\rho(\underline{R}_o,\underline{r}) = \rho(\theta_o) = ((\underline{R}-\underline{R}_o)^2 + z^2)^{\frac{1}{2}}$$

$$= (R^2 - 2RR_o\cos\theta_o + R_o^2 + z^2)^{\frac{1}{2}}$$

(dropping explicitly the other dependences of ρ). This is symmetric in θ_o, and $\rho(\theta_o) \geq z \geq 0$ for all $\underline{R}$ and $\underline{R}_o$. Then

$$H(R,z,\rho') = \frac{1}{2\pi}\int_0^\infty R_o\,dR_o\,G(R_o)\left[2\int_0^\pi d\theta_o\,\frac{\delta(\rho'-\rho(\theta_o))}{\rho(\theta_o)}\right]$$

Suppose $R \neq 0$ (otherwise the θ_o integral becomes trivial) and use

$$\delta(\rho'-\rho(\theta_o)) = \sum_i \frac{\delta(\theta_o-\theta_i)}{|\partial\rho/\partial\theta_o|}$$

where θ_i are solutions of $\rho' - \rho(\theta_i) = 0$. Then the integral I in square brackets becomes

$$I = 2\int_0^\pi d\theta_o \sum_i \frac{\delta(\theta_o-\theta_i)}{RR_o|\sin\theta_o|}$$

$$= \begin{cases} 2\sum_i \frac{1}{RR_o\sin\theta_i} & \text{if } \rho(\theta_i)=\rho' \text{ and } 0\leq\theta_i\leq\pi \\ 0 & \text{otherwise} \end{cases} \qquad (2.4)$$

Now $\rho' - \rho(\theta_i) = 0$

$$\Rightarrow \quad \rho'^2 = (\rho(\theta_i))^2 = R^2 - 2RR_o\cos\theta_i + R_o^2 + z^2$$

$$\Rightarrow \quad \rho'^2 - z^2 = R^2 - 2RR_o\cos\theta_i + R_o^2 = |\underline{R} - \underline{R}_o|^2.$$

Define $\rho_o = (\rho'^2 - z^2)^{\frac{1}{2}}$ = projection of ρ' onto the baffle plane.

Then $\rho_o^2 = (R^2 + R_o^2) - 2RR_o\cos\theta_i$

$$\Rightarrow \quad 2RR_o\sin\theta_i = ((2RR_o)^2 - (R^2 + R_o^2 - \rho_o^2)^2)^{\frac{1}{2}}$$

and the condition $\rho' = \rho(\theta_i)$ becomes $\rho_o^2 = |\underline{R} - \underline{R}_o|^2$ with $\rho' \geq z$.

This condition states that for a given ρ' and $\underline{r} \equiv (\underline{R},z)$, i.e. for given ρ_o and $\underline{R}$, the vector $\underline{R}_o$ must lie on the semicircle of radius ρ_o defined by $0 \leq \theta_i \leq \pi$ as in fig. 2.2.

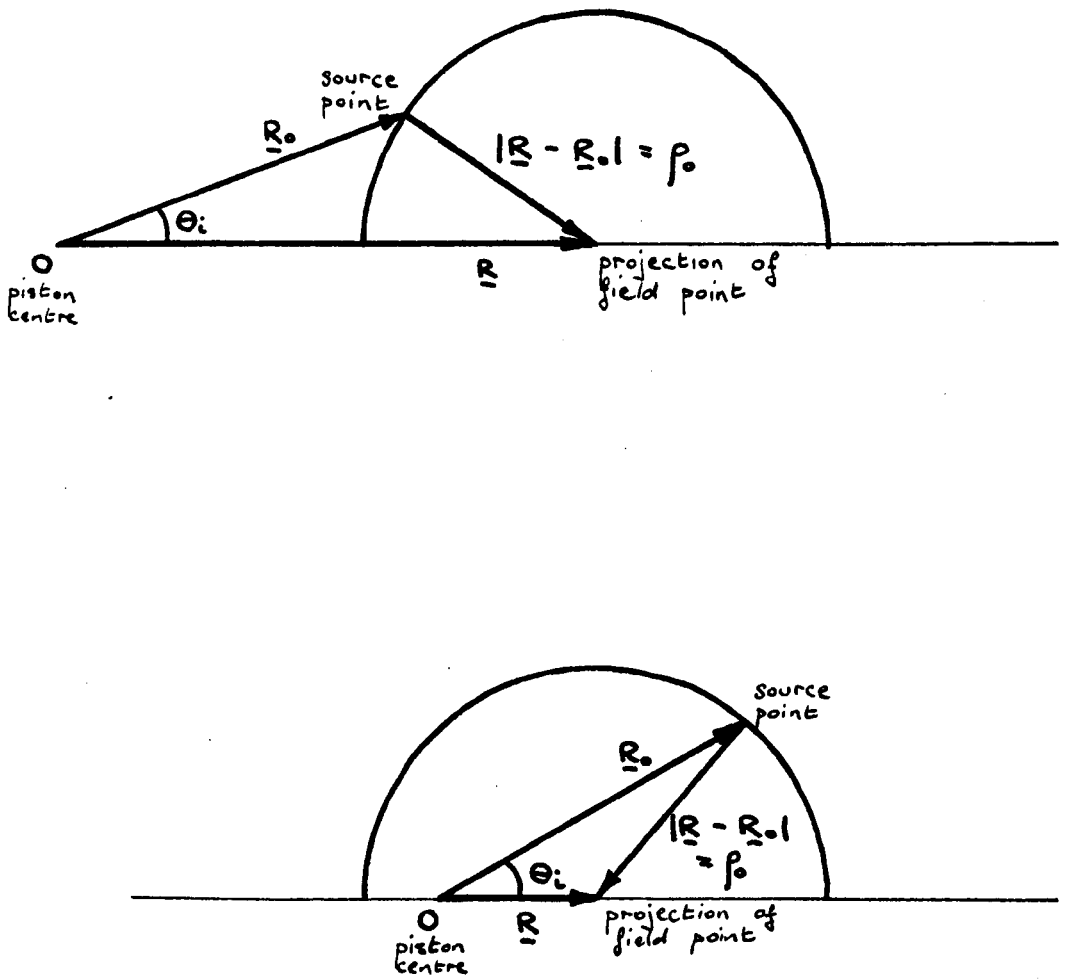


Figure 2.2 Configurations Contributing to I in equation (2.4)

If the triangle inequality $|R - \rho_0| \leq R_0 \leq R + \rho_0$ is satisfied, one real solution for θ_i can be found, otherwise no solution exists, so:

$$I = \begin{cases} \dfrac{4}{\sqrt{(2RR_0)^2 - (R^2+R_0^2-\rho_0^2)^2}} & \text{if } \rho' \geq z \text{ and } |R-\rho_0| \leq R_0 \leq R+\rho_0 \\ 0 & \text{otherwise} \end{cases}$$

We note that z only enters this expression through ρ_0, and that when either of the equalities $|R - \rho_0| = R_0$ (corresponding to $\theta_i = 0$) or $R_0 = R + \rho_0$ (corresponding to $\theta_i = \pi$) holds, then I diverges, as seen from (2.4), but the singularity is integrable.

Now we can write $H(R,z,\rho')$ in the form $H(R,z,\rho') = \mathcal{H}(R,\rho_0)\theta(\rho'-z)$ where $\rho_0 = (\rho'^2 - z^2)^{\frac{1}{2}}$ and

$$\mathcal{H}(R,\rho_0) = \frac{2}{\pi}\int_{|R-\rho_0|}^{R+\rho_0} \frac{G(R_0)\, R_0\, dR_0}{\sqrt{(2RR_0)^2 - (R^2+R_0^2-\rho_0^2)^2}} \qquad (2.5)$$

Then (2.2) becomes

$$\psi(R,z,t) = \int_{-\infty}^{\infty} d\rho'\, F'(t-\rho'/c)\,\mathcal{H}(R,\rho_0)\,\theta(\rho'-z)$$

Put $F'(t-\rho'/c) = -c\dfrac{\partial \rho_0}{\partial \rho'}\dfrac{\partial F}{\partial \rho_0}(t-\sqrt{\rho_0^2+z^2}/c)$

then $\psi(R,z,t) = -c\displaystyle\int_z^{\infty} d\rho' \frac{\partial \rho_0}{\partial \rho'}\frac{\partial F}{\partial \rho_0}\mathcal{H}(R,\rho_0)$

i.e. $$\psi(R,z,t) = -c\int_0^{\infty} d\rho_0 \frac{\partial F}{\partial \rho_0}(t-\sqrt{\rho_0^2+z^2}/c)\,\mathcal{H}(R,\rho_0) \qquad (2.6)$$

The integrand is the contribution to ψ of a circle on the baffle plane of radius ρ_0, centred on the foot of the perpendicular from the field point at $\underline{r}$.

We shall evaluate eqn.(2.5) for $\mathcal{H}(R,\rho_0)$ further in section 2.4, but now we shall pause to consider our current formula for the wavefunction, equation (2.6).

2.3 Discussion of the General Formulation

The physical significance of (2.6) becomes more apparent if it is integrated by parts, but first we need more information about $\mathcal{H}(R,\rho_0)$. We wish to evaluate $\mathcal{H}(R,\rho_0)$ as $R \to 0$, $\rho_0 \to 0$, $R \to \infty$, and $\rho_0 \to \infty$. In each case the range of integration tends to zero relative to the position of the centre of the range, but the integrand tends to infinity. The effect is to factor out a particular value of G (assuming G is continuous at these points) leaving the integral

$$\frac{2}{\pi}\int_{|R-\rho_0|}^{R+\rho_0} \frac{R_0\,dR_0}{\sqrt{(2RR_0)^2 - (R^2+R_0^2-\rho_0^2)^2}} = 1 \qquad (2.7)$$

(evaluated by putting $x = R_0^2 - R^2 - \rho_0^2$). We find that

$$\mathcal{H}(0,\rho_0) = G(\rho_0)$$
$$\mathcal{H}(R,0) = G(R)$$
$$\mathcal{H}(R,\infty) = G(\infty)$$
$$\mathcal{H}(\infty,\rho_0) = G(\infty).$$

These limits can also be deduced directly from the definition of $\mathcal{H}(R,\rho_0)$.

Now integrating (2.6) by parts and assuming $G(\infty) = 0$ for any real radiator, or $F(-\infty) = 0$ for any real pulse

$$\psi(R,z,t) = cF(t-z/c)\,G(R) + c\int_0^\infty d\rho_0\, F(t-\sqrt{\rho_0^2+z^2}/c)\,\frac{\partial \mathcal{H}}{\partial \rho_0}(R,\rho_0) \qquad (2.8)$$

To understand this expression, suppose the whole baffle plane vibrates rigidly so that $G(R_0) = 1$ everywhere. Then we can take this outside the integral for $\mathcal{H}(R,\rho_0)$, which becomes (2.7), giving $\mathcal{H}(R,\rho_0) = 1$ and $\partial\mathcal{H}/\partial\rho_0 = 0$. Then the second term in (2.8) is zero, and $\psi(R,z,t) = cF(t-z/c)$. The fluid velocity in the wave is then

$$-\nabla\psi = -c\,\frac{\partial F}{\partial z}(t-z/c)\,\hat{\underline{z}}$$

$$= F'(t-z/c)\,\hat{\underline{z}}$$

as expected, because the velocity of the baffle plane is $F'(t)\,\hat{\underline{z}}$. Hence the first term in (2.8) represents a plane wave pulse emitted by the point on the radiator directly below the field point, and the second term represents the contribution by the rest of the radiator. (Compare (2.8) with a similar result for the CW case due to Schoch(1941) as discussed in Carter & Williams(1951) and Williams(1951)).

If $G(\underline{R}_o)$, and therefore $\mathcal{H}$, is slowly varying across the radiator the second integral will be small, and the largest contributions will come from discontinuities of $\mathcal{H}$ such as arise from the edges of a piston radiator (see later). These discontinuities produce a sequence of "replica pulses" which are distorted and delayed versions of the initial pulse represented by the first term. These were first discussed by Freedman(1970), who also used a "geometry function" similar to ours. Stepanishen (1971c) has shown that they produce interesting transient phenomena at the CW nulls of a circular piston radiator. Beaver (1974) displays some computed wavefunctions showing the two replica pulses produced by a rigid circular piston radiator, and he states that "... the discontinuity effects of the rims eventually overlap the main pulse, producing interference. The pulse amplitude is then more similar to the CW case. However, distinct nulls cannot form because complete interference is not possible". We shall see later that this is not true: although nulls fixed in space are highly unlikely with pulses, moving nulls, i.e. dislocations, certainly occur! Note that if R is outside the radiating area then G(R) is zero, and the initial pulse is missing leaving only the replicas.

We can demonstrate the replica pulses by the simple case of an axial field point (R = 0) for a rigid circular piston radiator of radius a, such that $G(R_o) = \theta(a - R_o)$. Then

$$\mathcal{H}(0,\rho_o) = G(\rho_o) = \theta(a - \rho_o)$$

so $$\frac{\partial \mathcal{H}}{\partial \rho_o}(0,\rho_o) = -\delta(a - \rho_o)$$

and $$\psi(0,z,t) = c\{F(t-z/c) - F(t-\sqrt{z^2+a^2}/c)\} \qquad (2.9)$$

This is a well known result, usually quoted only for the CW case (e.g. see Rschevkin (1963), p434, eqn.11-21). In this high symmetry position there is just one replica pulse, which is an inverted (but not distorted) and delayed version of the initial pulse: it is the "edge wave" from the rim of the piston. Notice that the velocity potential wavefunction is simply related to the radiator displacement, and hence the pressure wavefunction is simply related to the radiator velocity, in the near field.

2.4 The Rigid Circular Piston Radiator

In this section we specialize the formula for $\mathcal{H}(R,\rho_o)$ to the case of a rigid plane circular piston vibrating in an infinite plane baffle, as shown in fig.2.1, such that $G(R_o) = \theta(a - R_o)$. From (2.5) we have

$$\mathcal{H}(R,\rho_o) = \frac{2}{\pi}\int_{|R-\rho_o|}^{R+\rho_o} \frac{\theta(a-R_o)\,R_o\,dR_o}{\sqrt{(2RR_o)^2 - (R^2+R_o^2-\rho_o^2)^2}} = \frac{2}{\pi}\int_{|R-\rho_o|}^{\alpha} \frac{R_o\,dR_o}{\sqrt{\quad}}$$

where
$$\alpha = \begin{cases} |R-\rho_o| & \text{if } a < |R-\rho_o| \quad \text{(I)} \\ a & \text{if } |R-\rho_o| \leq a \leq R+\rho_o \quad \text{(II)} \\ R+\rho_o & \text{if } R+\rho_o < a \quad \text{(III)} \end{cases}$$

We evaluate the integral by putting $x = R_o^2 - R^2 - \rho_o^2$ giving

$$\frac{1}{\pi}\int_{-2R\rho_o}^{\alpha^2-R^2-\rho_o^2} \frac{dx}{\sqrt{(2R\rho_o)^2 - x^2}}$$

$$= \frac{1}{\pi}\left(\arcsin\frac{\alpha^2-R^2-\rho_o^2}{2R\rho_o} + \arcsin 1\right)$$

In regions

(I) : $\mathcal{H} = \frac{1}{\pi}(-\arcsin 1 + \arcsin 1) = 0$

(II) : $\mathcal{H} = \frac{1}{\pi}\arcsin\frac{a^2-R^2-\rho_o^2}{2R\rho_o} + \frac{1}{2}$

(III) : $\mathcal{H} = \frac{1}{\pi}(\arcsin 1 + \arcsin 1) = 1$

Rewriting the conditions in terms of ρ_o we have the following three situations, illustrated in fig.2.3:

(I) $0 \leq \rho_o < R - a$ (if $R > a$) and $\rho_o > R + a$

$\mathcal{H}(R,\rho_o) = 0$ because the source line S is outside the piston P.

(II) $|R - a| \leq \rho_o \leq R + a$

$\mathcal{H}(R,\rho_o) = \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{a^2-R^2-\rho_o^2}{2R\rho_o}$ because S is partly inside P.

(III) $0 \leq \rho_o < a - R$ (if $R < a$)

$\mathcal{H}(R,\rho_o) = 1$ (as for an infinite plane radiator) because S is completely inside P.

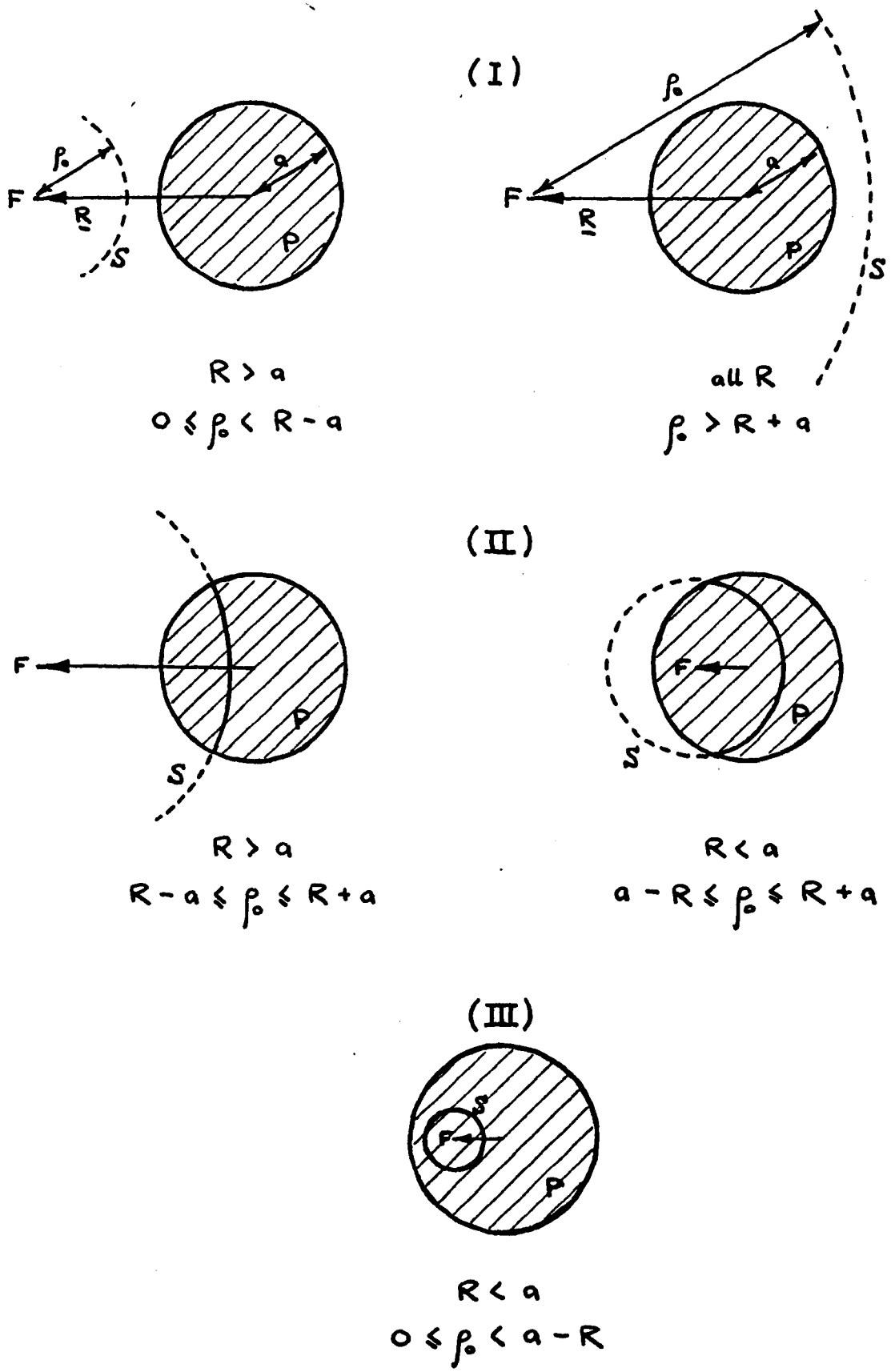


Figure 2.3 The Three Forms taken by the Source Line S
F is the projection of the field point
P is the piston face (hatched)
The emitting part of S is shown solid

We distinguish three different field point regions, which give the different forms for $\mathcal{H}(R,\rho_o)$ sketched in fig.2.4:

1) $\underline{R < a}$ - the field point is "inside" the piston. As ρ_o increases from 0 the source line S is entirely within the piston P, and $\mathcal{H}$ has the plane wave value of 1, until at $\rho_o = a - R$ S touches the edge of P and $\partial\mathcal{H}/\partial\rho_o$ jumps discontinuously from 0 to $-\infty$. $\mathcal{H}$ then decreases monotonically as progressively less of S remains in P, until at $\rho_o = a + R$ S touches the other edge of P and $\partial\mathcal{H}/\partial\rho_o$ again jumps discontinuously from $-\infty$ to 0. After this $\mathcal{H} = 0$ because S is outside P.

2) $\underline{R > a}$ - the field point is "outside" the piston. As ρ_o increases from 0, $\mathcal{H} = 0$ because S has not yet reached the piston. At $\rho_o = R - a$ S touches the edge of the piston and $\partial\mathcal{H}/\partial\rho_o$ jumps from 0 to $+\infty$. $\mathcal{H}$ then increases to a maximum, which is always less than half the plane wave value, at which the length of S inside P is a maximum. It then decreases to 0 at $\rho_o = R + a$, where $\partial\mathcal{H}/\partial\rho_o$ jumps from $-\infty$ to 0 again as S touches the other edge of P. After this $\mathcal{H} = 0$ because S is "outside" the piston.

3) $\underline{R = a}$ - the field point is "on the edge" of the piston. This is the limit of both the above cases as $R \to a$. As ρ_o increases $\mathcal{H}$ decreases monotonically from half the plane wave value.

(c.f. graphs of the impulse response functions shown by Lochwood & Willette (1973))

If we define

$$\mathcal{G}(R,\rho_o) = \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{a^2-R^2-\rho_o^2}{2R\rho_o}$$

then $\mathcal{H}(R,\rho_o) = \theta(a-R-\rho_o) + \theta(\rho_o-|R-a|)\theta((R+a)-\rho_o)\,\mathcal{G}(R,\rho_o)$

(in the sense of generalised functions). Note that

$$\mathcal{G}(R,R+a) = 0$$

$$\mathcal{G}(R,|R-a|) = \theta(a-R) \qquad R \neq a$$

$$\mathcal{G}(a,0) = \tfrac{1}{2}$$

as shown on the graphs in fig.2.4.

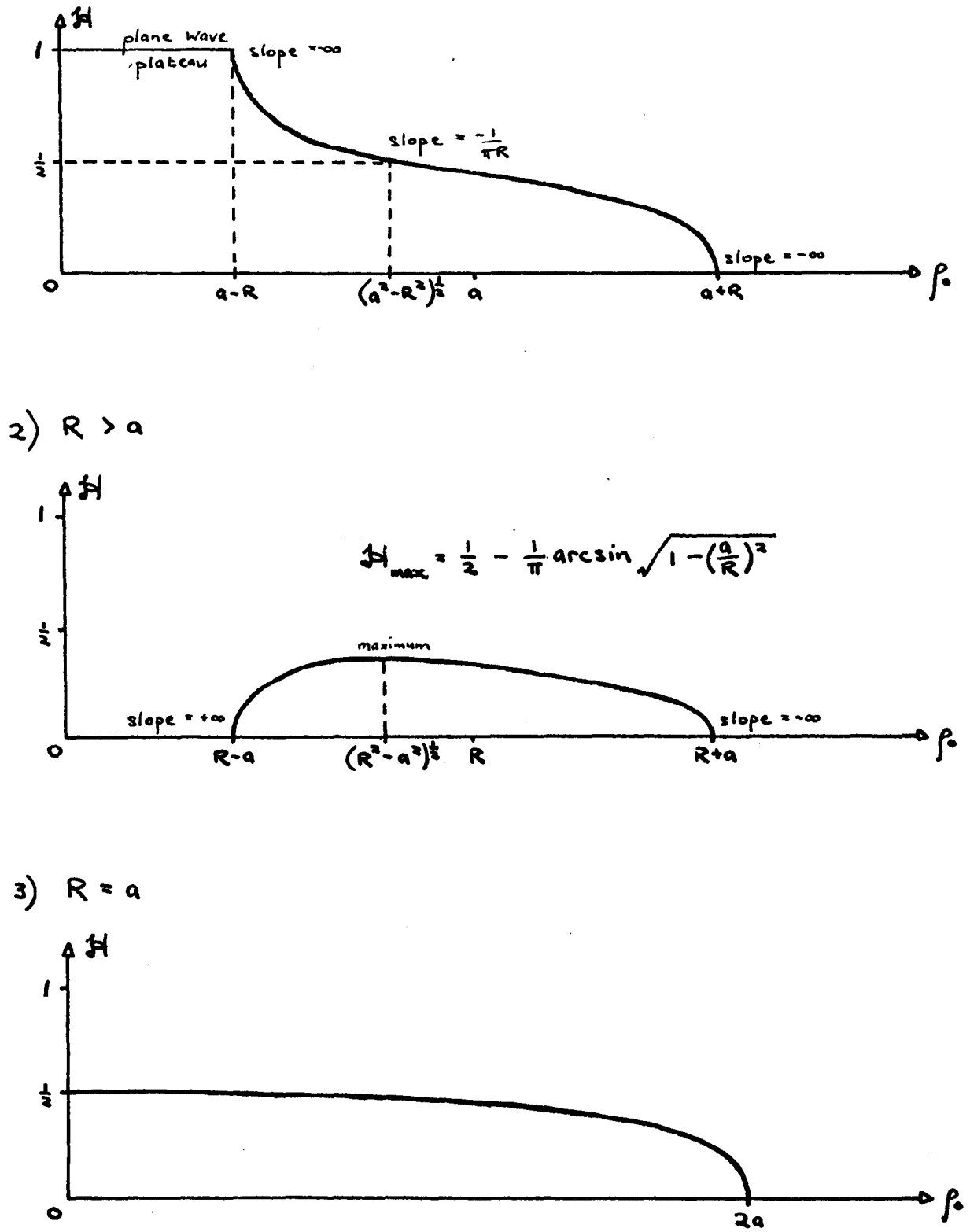


Figure 2.4 Geometry function $\mathcal{H}$ for the three different Field Point Regions

If we define $\Theta(x) = 0$ if $x < 0$

$\frac{1}{2}$ if $x = 0$

1 if $x > 0$

then $\mathcal{G}(R,|R-a|) = \Theta(a-R)$ for all R.

To avoid having to differentiate generalised functions (which becomes tricky at R = a) we use eqn.(2.6) rather than (2.8), and substitute the above expression for $\mathcal{H}(R,\rho_o)$ to give

$$\psi(R,z,t) = -c\int_0^\infty d\rho_o \frac{\partial F}{\partial \rho_o}(t - \sqrt{\rho_o^2+z^2}/c)\mathcal{H}(R,\rho_o)$$

$$= -c\Theta(a-R)\int_0^{a-R} d\rho_o \frac{\partial F}{\partial \rho_o} - c\int_{|R-a|}^{R+a} d\rho_o \frac{\partial F}{\partial \rho_o}\mathcal{G}(R,\rho_o)$$

i.e. $$\psi(R,z,t) = \Theta(a-R)c\{F(t - z/c) - F(t - \sqrt{z^2+(a-R)^2}/c)\} \quad (1)$$
$$-c\int_{|R-a|}^{R+a} d\rho_o \frac{\partial F}{\partial \rho_o}\mathcal{G}(R,\rho_o) \quad (2) \qquad (2.10)$$

Reference to eqn.(2.9) shows that the term (1) in braces is the axial field of a circular piston of radius (a-R). If the field point is "inside" the piston (i.e. R < a), this term gives the contribution of the largest circular region of the piston centred on the projection of the field point, and the integral (2) gives the contribution of the rest (see fig.2.5). If the field point is "outside" the piston the first term does not contribute. At R = a the term in braces is zero so there is no discontinuity.

Now we can conveniently integrate (2.10) by parts, when the integral (2) becomes

$$cF(t - \sqrt{(R-a)^2+z^2}/c)\Theta(a-R) + c\int_{|R-a|}^{R+a} d\rho_o \, F \frac{\partial \mathcal{G}}{\partial \rho_o}(R,\rho_o)$$

giving

$$\psi(R,z,t) = \Theta(a-R)cF(t-z/c) + c\int_{|R-a|}^{R+a} d\rho_o \, F \frac{\partial \mathcal{G}}{\partial \rho_o}(R,\rho_o) \qquad (2.11)$$

which is eqn.(2.8) specialized to the rigid circular piston.

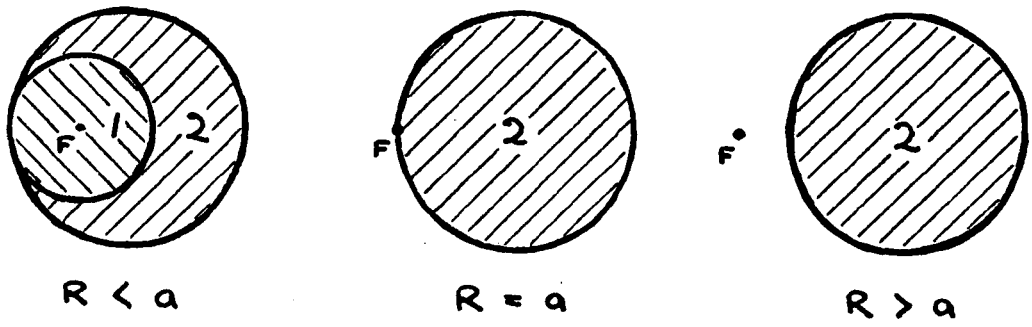


Figure 2.5 Contributions of Terms of equation (2.10)

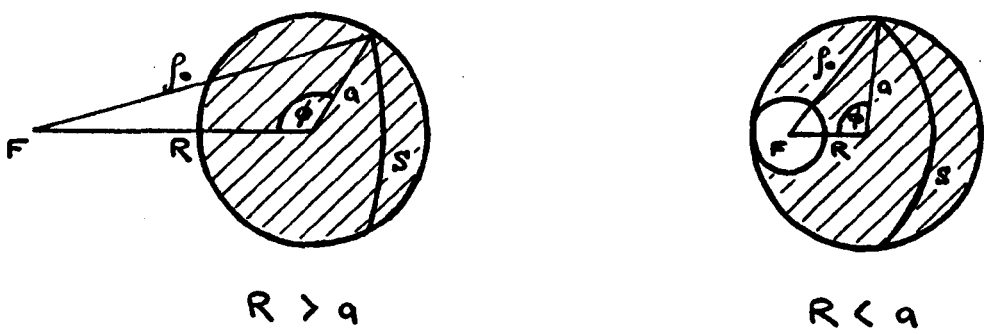


Figure 2.6 Significance of Angle ϕ in equation (2.12)

The integral term is

$$\frac{c}{\pi}\int_{|R-a|}^{R+a} \frac{d\rho_o}{\rho_o} \frac{R^2-a^2-\rho_o^2}{\sqrt{(2R\rho_o)^2-(a^2-R^2-\rho_o^2)^2}} F(t - \sqrt{z^2+\rho_o^2}/c)$$

where the integrand diverges at both ends of the range. To avoid this, we define a new variable of integration ϕ by $\rho_o^2 = a^2 + R^2 - 2aR\cos\phi$ giving

$$\frac{ac}{\pi}\int_0^{\pi} d\phi \left(\frac{R\cos\phi - a}{a^2+R^2-2aR\cos\phi}\right) F(t - \sqrt{z^2+a^2+R^2-2aR\cos\phi}/c)$$

This integrand is regular everywhere, and has proved convenient for computation.

Physically, the integrand is the contribution of a curved source line S whose position on the piston is parametrised by ϕ, and as ϕ varies from 0 to π this source line moves over the whole area of the piston, excluding the region (if any) where it would be a complete circle (see fig.2.6). Our main result, then, is that for <u>any</u> piston <u>displacement</u> pulse F(t) the wavefunction at <u>any</u> point is <u>exactly</u>:

$$\psi(R,z,t) = \Theta(a-R)cF(t - z/c) + \frac{ac}{\pi}\int_0^{\pi} d\phi \left(\frac{R\cos\phi - a}{a^2+R^2-2aR\cos\phi}\right) F(t - \sqrt{z^2+a^2+R^2-2aR\cos\phi}/c) \quad (2.12)$$

(This can easily be deduced geometrically for the case R = a, and it can probably be done generally).

Note that the acoustic pressure $\delta p = d\,\partial\psi/\partial t$ and that $\partial\psi/\partial t$ is obtained from eqn.(2.12) by replacing F by $\partial F/\partial t$. Then interpreting F(t) as the piston velocity, instead of displacement, gives $\delta p = d\,\psi(R,z,t)$. This allows us to compare our CW results with those of other authors.

2.5 The Far Field (Fraunhofer) Limit

Let us change to spherical polar coordinates (r,θ) such that $R = r\sin\theta$ and $z = r\cos\theta$. Then the far field limit is the limiting form as $r \to \infty$. Suppose $\theta \neq 0$ $(\Rightarrow R \neq 0)$, then (2.12) becomes

$$\frac{ac}{\pi}\int_0^{\pi} d\phi \left(\frac{r\sin\theta\cos\phi - a}{a^2+r^2\sin^2\theta-2ar\sin\theta\cos\phi}\right)$$

$$\times F(t - \sqrt{r^2\cos^2\theta+a^2+r^2\sin^2\theta-2ar\sin\theta\cos\phi}/c)$$

Taking the limit $r \to \infty$ this is

$$\frac{ac}{\pi}\int_0^{\pi} d\phi \left(\frac{\cos\phi}{r\sin\theta}\right) F(t - (r-a\sin\theta\cos\phi+O(1/r))/c) + O(1/r^2)$$

Working to lowest order, we define the far field limit as

$$\psi_F(r,\theta,t) = \frac{ac}{\pi r\sin\theta}\int_0^{\pi} d\phi\ \cos\phi\ F(t - \frac{r}{c} + \frac{a}{c}\sin\theta\cos\phi) \qquad (2.13)$$

Physically this simplification occurs because the curved source line responsible for eqn.(2.12) has become straight, as in the conventional far field calculation.

On the axis $R = 0$, this formula is indeterminate, so we take the limit $\theta \to 0$ using L'Hospital's rule to give

$$\psi_F(r,0,t) = \frac{a^2}{2r} F'(t - r/c) \qquad (2.14)$$

But eqn.(2.12) takes a different form from that above for $R = 0$, viz:

$$\psi(0,z,t) = c\{F(t - z/c) - F(t - \sqrt{z^2+a^2}/c)$$

as deduced before (eqn.(2.9)). We require the limiting form as $z \to \infty$:

$$\psi(0,z,t) \sim c\{F(t - z/c) - F(t - (z+a^2/2z)/c)\}$$

$$\sim \frac{a^2}{2z} F'(t - z/c)$$

which agrees with (2.14).

Physically, the replica pulse from the edge of the piston is combining with the initial pulse, with identical amplitude and opposite sign, so as to differentiate the initial pulse in the far field. So whilst in the nearfield the velocity potential wavefunction is simply related to the piston displacement, in the farfield it is simply

related to the piston velocity. At this point it must be remembered that we have assumed F(t) to be differentiable at least once, otherwise the results would not strictly apply. We illustrate this differentiation effect in fig.2.7, for a trapezoidal pulse, where the variables used are $Z = z/a$, $T = ct/a$ and the function $f(T)$ is defined by $cF(t) = f(ct/a)$. Then the axial wavefunction is

$$f(T - Z) - f(T - \sqrt{Z^2+1}) \cdot$$

$$\to \frac{1}{2Z} f'(T - Z) \quad \text{as} \quad Z \to \infty.$$

We plot the driving pulse $f(T)$, the axial wavefunction at $Z = 0,1,2,3,4,5$ and the farfield form for $Z = 5$. We cannot expect the farfield form to be very accurate so close to the piston, but in fact the only significant error is in the pulse shape. This is because our driving pulse is not strictly differentiable (or very realistic), and a smoother pulse such as a Gaussian would give much better agreement. Then the positive and negative pulses would join smoothly, with a single zero in the middle, everywhere along the axis, despite the fact that the driving pulse may have no zeros at finite times. We seem to have produced something akin to a dislocation running along the axis. It would be interesting to study this type of non-quasimonochromatic pulse more generally, but this has not yet been done.

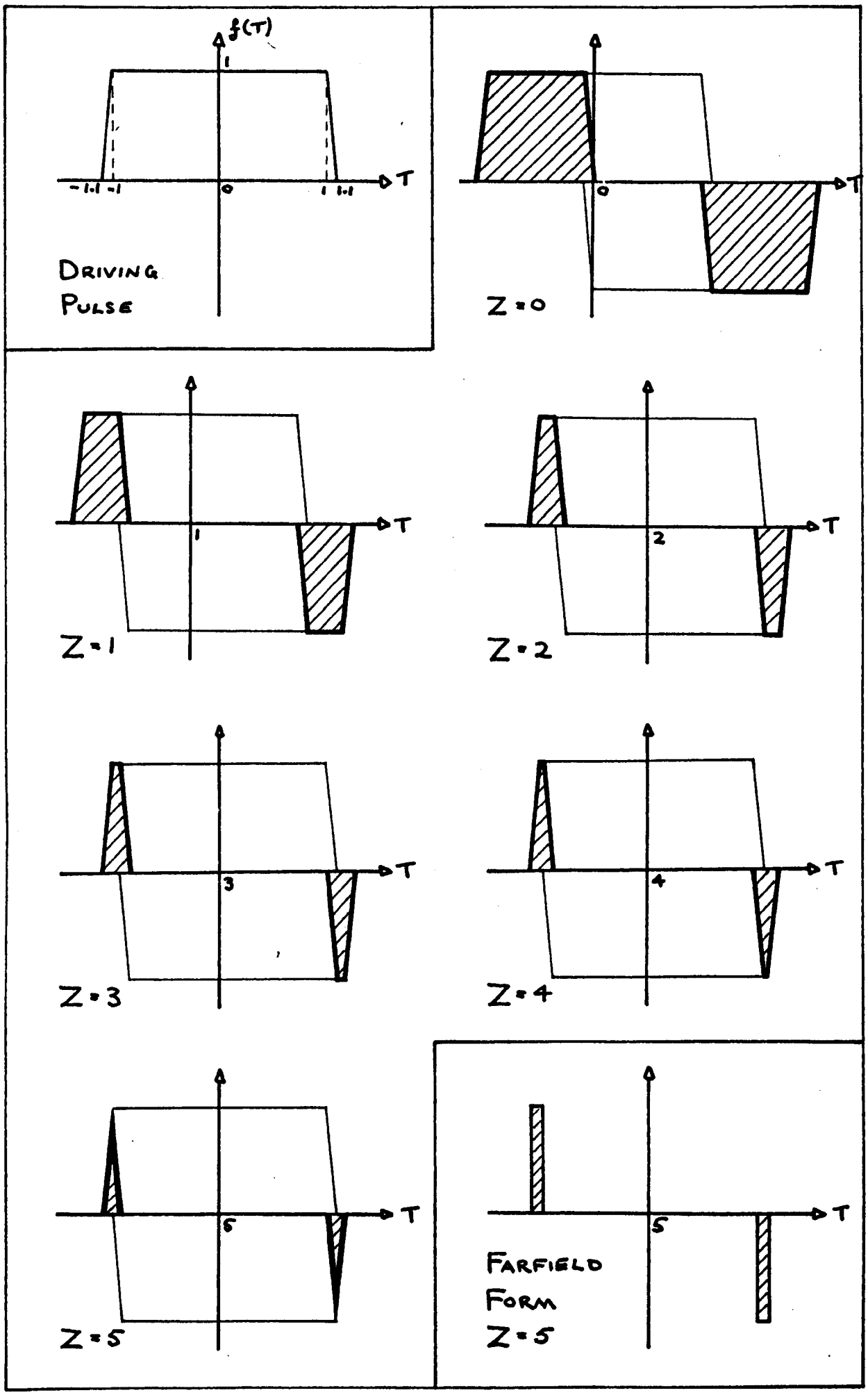


Figure 2.7 Axial Approach to the Far Field

2.6 Continuous Waves and their Nulls

Let us drive the piston with the monochromatic signal

$$F(t) = e^{i\omega t} \qquad \text{(strictly taking the real or imaginary part)}$$

and define the dimensionless variables

$$\mathcal{R} = R/a, \quad Z = z/a, \quad T = \omega t, \quad K = ka = \omega a/c \quad \text{and} \quad \Psi(\mathcal{R},Z,T) = \frac{1}{c}\psi(R,z,t).$$

Then the monochromatic wavefunction is

$$\Psi_m(\mathcal{R},Z,T) = \Theta(1-\mathcal{R})e^{i(T-KZ)} + \frac{1}{\pi}\int_0^{\pi} d\phi \left(\frac{\mathcal{R}\cos\phi - 1}{1+\mathcal{R}^2-2\mathcal{R}\cos\phi}\right) e^{i(T - K\sqrt{Z^2+1+\mathcal{R}^2-2\mathcal{R}\cos\phi})} \qquad (2.15)$$

remembering that for $\mathcal{R} = 1$ we have

$$\left(\frac{\mathcal{R}\cos\phi - 1}{1+\mathcal{R}^2-2\mathcal{R}\cos\phi}\right) = -\frac{1}{2}$$

to avoid the indeterminacy at $\phi = 0$.

On the axis we use the simpler form

$$\Psi_m(0,Z,T) = e^{i(T - KZ)} - e^{i(T - K\sqrt{Z^2+1})} \qquad (2.16)$$

In the far field we have from (2.13)

$$\psi_{Fm}(r,\theta,t) = e^{i(\omega t-kr)} \frac{ac}{\pi r \sin\theta}\int_0^{\pi} d\phi \cos\phi\, e^{i(K\sin\theta)\cos\phi}$$

$$= ie^{i(\omega t-kr)} \frac{ac}{r\sin\theta} J_1(K\sin\theta) \qquad (2.17)$$

where $J_1(x)$ is a Bessel function (see e.g. Bell (1968), p102). This is a standard result (usually calculated for $F'(t) = e^{i\omega t}$, e.g. see Rschevkin (1963), eqn.11-28) showing that the maximum far field amplitude occurs on the axis, where (2.14) or (2.17) gives

$$\psi_{Fm}(r,0,t) = i\omega e^{i(\omega t-kr)}\frac{a^2}{2r}.$$

Therefore, let us normalize the far field wavefunction by dividing by the maximum amplitude $\omega a^2/2r$, and define a dimensionless retarded time $\tau = \omega t - kr$ and a reduced farfield wavefunction

$$\Psi_{Fm}(\theta,\tau) = \frac{2r}{\omega a^2}\psi_{Fm}(r,\theta,t) = ie^{i\tau}\frac{2J_1(K\sin\theta)}{K\sin\theta} \qquad (2.18)$$

From eqn.(2.18) we see that farfield (angular) nulls occur where $J_1(K\sin\theta) = 0 \Rightarrow K\sin\theta = j_{1n}$ where j_{1n} is the nth zero of J_1, and asymptotically has the value $(n + 1/4)\pi$ (e.g. see Bell (1968), p127). The first few values are (from Abromowitz & Stegun (1965)):

n	j_{1n}	$(n+1/4)\pi$
1	3·8317	3·927
2	7·0156	7·069
3	10·1735	10·210
4	13·3237	13·352
5	16·4706	16·493

If $j_{1n} < K \leq j_{1,n+1}$ there are n farfield nulls, i.e. there are approximately $\text{Int}(K/\pi - 1/4)$ farfield nulls ($\text{Int}(x)$ = maximum integer $\leq x$).

From eqn.(2.16) the axial CW wavefunction is

$$\Psi_m(0,Z,T) = e^{iT}\left(e^{-iKZ} - e^{-iKZ_1}\right) \quad \text{where } Z_1 = \sqrt{Z^2+1}$$

$$= 2ie^{i(T - K\tilde{Z})}\sin K\left(\frac{Z_1-Z}{2}\right) \tag{2.19}$$

where $\tilde{Z} = \frac{Z+\sqrt{Z^2+1}}{2}$ = mean distance from centre and edge of piston.

Axial nulls occur when $K\left(\frac{Z_1-Z}{2}\right) = n\pi$ (2.20)

where n is a natural number, i.e. at $Z_n = \frac{K^2-(2n\pi)^2}{4n\pi K}$ which must be $\geqslant 0$, therefore there are exactly $\text{Int}(K/2\pi)$ axial nulls. In fact, since $K/2\pi = a/\lambda$, the number of axial nulls is the number of whole wavelengths which fit into the piston radius!

There seems to be no correlation between the number of far field nulls and the number of axial nulls, other than that for large K there are approximately twice as many farfield nulls as axial nulls. The relation between CW nulls will be seen to be important later for pulses, and is clearly not simple.

For computation we use $K = 10.0$ ($\Rightarrow \lambda/a = 0.63$), partly to model some experimental work being done in the laboratory using ultrasonic pulses in air (in which $c = 331.5\times10^3$ cm/s. For a = 0.5 cm, a frequency of 100 KHz gives $\lambda = 0.33$ cm). This value gives <u>one axial null</u> at $Z = 0.4816$, and <u>two farfield nulls</u> at

$$10 \sin\theta = 3.8317 \quad \Rightarrow \quad \theta = 22.53^{\circ}$$
$$7.0156 \qquad\qquad\quad 44.55^{\circ}$$

Note that for a slightly larger K a new farfield null would appear at $\theta = 90^{\circ}$ (without immediately introducing a new axial null), and hence we can expect to see vestiges of this incipient null in the diffraction patterns.

Fig.2.8a shows the reduced farfield amplitude, and fig.2.8b the phase (for $\tau = 0$) as a function of θ, from eqn.(2.18). The phase jump of π at a null produces the comb-like pattern of equiphase lines (fig. 2.8c) discussed in section 1.1 and illustrated in fig.1.3 of chapter 1. We can progress no further analytically, and must evaluate our wavefunctions numerically from now on. The integrals can all be done easily using standard routines (those using Chebyschev series were found to be very satisfactory) despite the oscillatory integrands. The contour plots were produced using a standard routine, but the three-dimensional plots were produced by a specially written routine. The ticks (solidi) indicate where the graphs cross zero.

The nearfield continuous wavefunction is computed from eqn.(2.15) and eqn.(2.16). Figs.2.9 and 2.10a show the nearfield amplitude. There is only one null, which is on the axis at $Z = 0.48$. The phase singularity shows up clearly on fig.2.10b. The amplitude has valleys, whose bottoms are generally not at zero height, running off into the far field. One valley begins at the axial null, runs up to a saddle point at about $Z = 0.9$, $R = 0.3$, and then runs down into the far field with monotonically decreasing height, which becomes zero at infinity. The other main valley begins at a minimum just in front of the piston, near $Z = 0.1$, $R = 0.4$, runs up to a saddle point near $Z = 0.25$, $R = 0.6$, and then down into the far field. The two minima are linked by a short valley passing over the saddle near $Z = 0.25$, $R = 0.25$. The curvature of the contours close to the piston near $R = 1$ in fig.2.10a suggests another slight valley, and fig.2.9 clearly shows a slight dip here: this is the vestige of the next farfield null (and associated valley) to appear as K is increased. Figs.2.10a & b show the farfield null angles as dashed lines. Note that fig.2.10b shows the equiphase lines bunching together in the minima near the piston, tending towards the coalescence which occurs at an actual null.

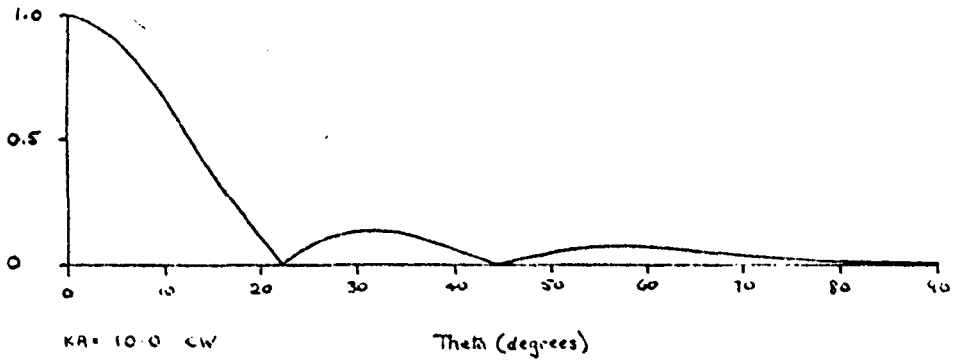


Figure 2.8a
REDUCED FARFIELD AMPLITUDE

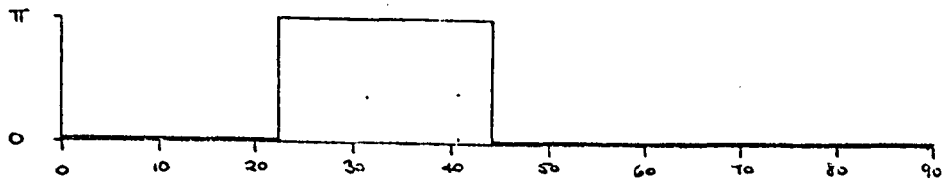


Figure 2.8b
PHASE - π/2

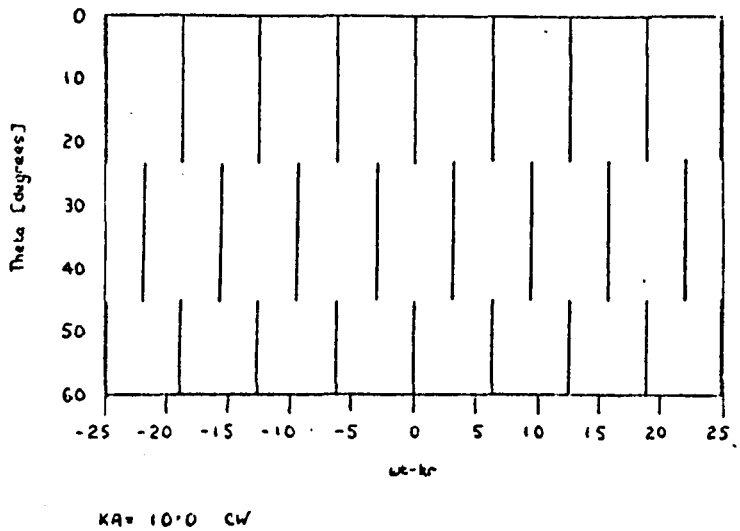


Figure 2.8c
REDUCED FARFIELD WAVEFN
CONTOURS OF PHASE [π/2]

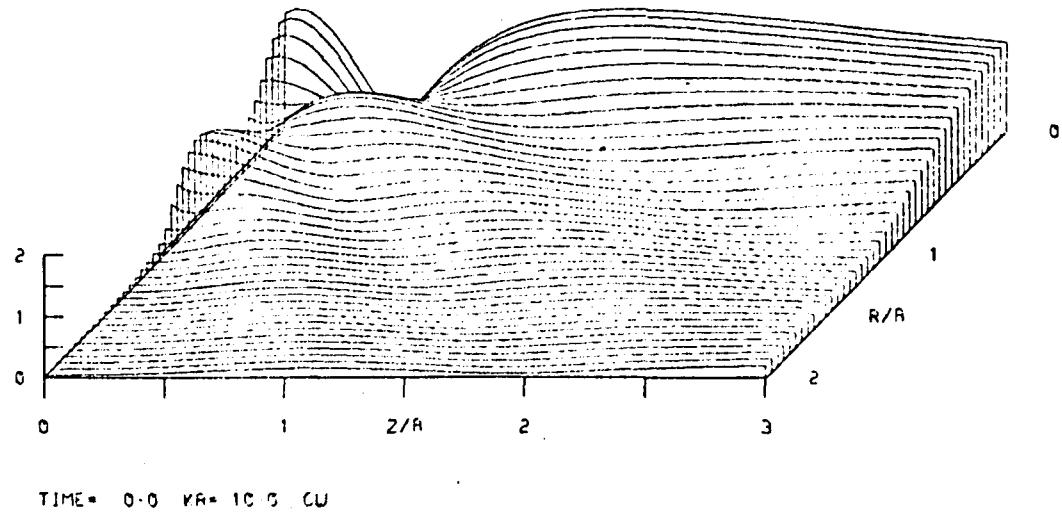


Figure 2.9
NEARFIELD WAVEFUNCTION
AMPLITUDE

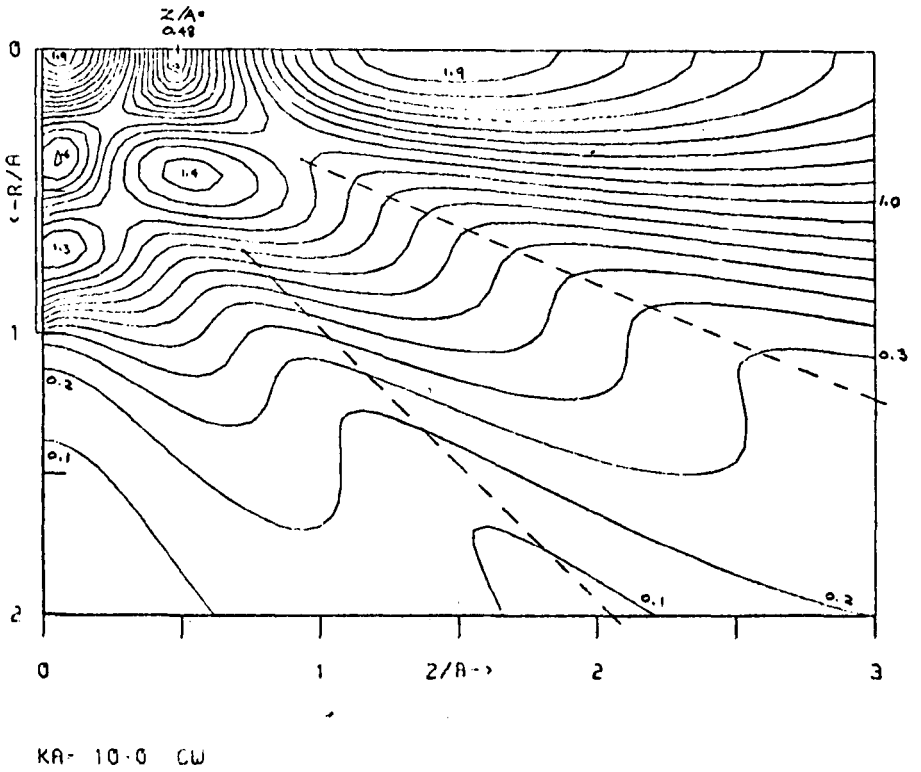


Figure 2.10a
NEARFIELD WAVEFUNCTION
CONTOURS OF AMPLITUDE in steps of 0.1

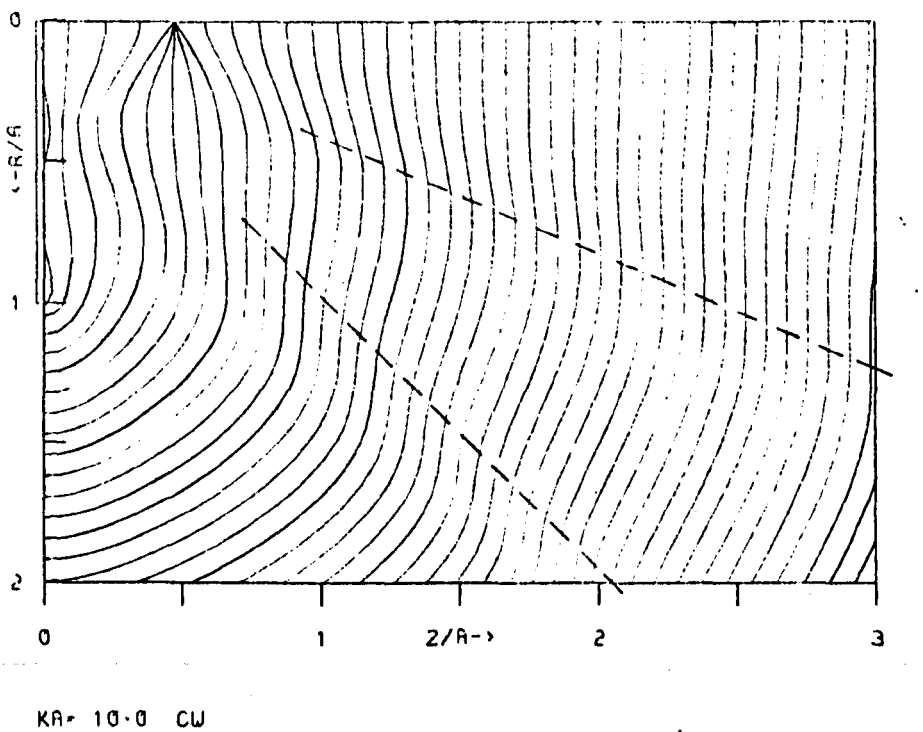


Figure 2.10b
NEARFIELD WAVEFUNCTION
CONTOURS OF PHASE in steps of Pi/4

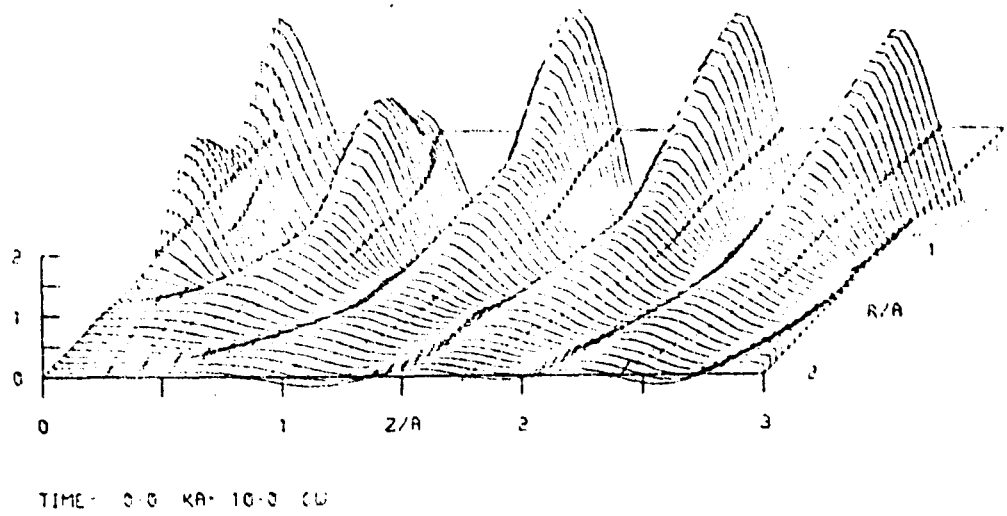
Figs.2.11a & b show the actual wavefunction, one with a phase shift of $\pi/2$ relative to the other. These, and fig.2.9, show what looks like approximately standing waves across the piston face. Every point over the piston face is nearly in phase (see fig.2.10b), but the wavelength seems to be somewhat longer than the real wavelength. Fig.2.11b shows particularly clearly the effect of the axial null, where a wavecrest comes to an abrupt end. Bearing in mind that the graphs should be rotated about the axis to generate the full wavefield, the effect is that the wavefronts develop a puncture as they pass through the null, which immediately closes again on the other side. The point puncture is fixed in space as the wavefronts sweep through it.

The complicated diffraction effects seem to be contained within a sphere of radius about 1.5a around the centre of the piston. Outside this radius the amplitude is decreasing monotonically, and the wavefronts (or equiphase lines) look like distorted spherical waves. This distortion takes the form of kinks near the farfield null angles, which are tending towards the farfield discontinuities. The standard definition of the nearfield/farfield "boundary" is the axial maximum of amplitude furthest from the radiator, which occurs at

$$Z = \frac{K^2 - \pi^2}{2\pi K} = 1.4345 \quad \text{for} \quad K = 10.$$

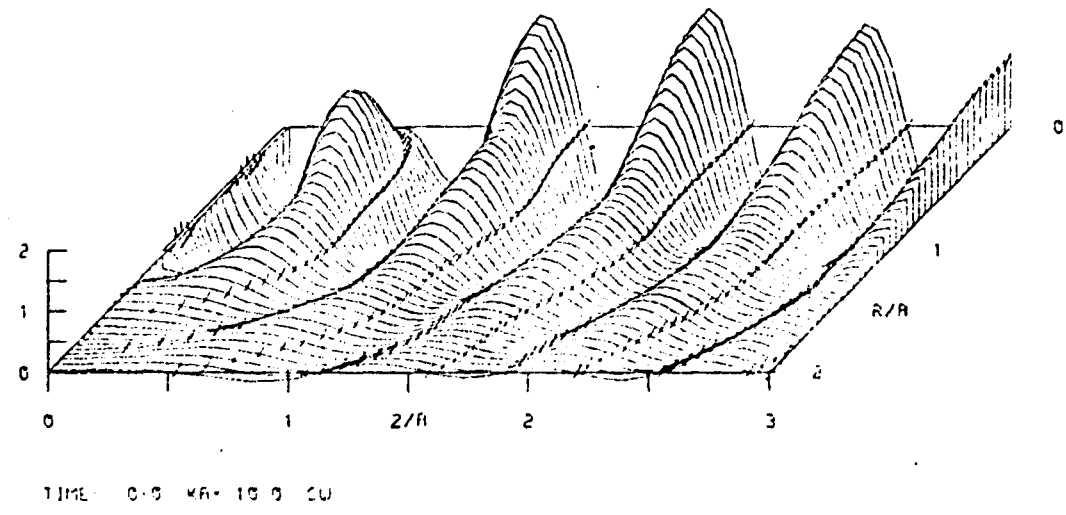
This is also the point where the paths from the centre and the edge of the piston differ by exactly $\lambda/2$ i.e. when the first Fresnel zone just fills the piston face. The graphs show that outside this boundary the wavefunction is farfield-like, except that the amplitude minima are not zeros, thereby supporting this definition of the boundary.

Our plots of CW amplitude agree quite well with Stenzel's plot shown on p443 of Rschevkin (1963), except that Stenzel shows no evidence of the incipient null. They also agree qualitatively with plots for larger values of a/λ displayed by Zemanek(1971) and Lockwood & Willette (1973). No-one has previously displayed the equiphase lines, or detailed plots of the actual wavefunctions.



TIME= 0.0 KA= 10.0 CU

Figure 2.11a
NEARFIELD WAVEFUNCTION
REAL PART



TIME= 0.0 KA= 10.0 CU

Figure 2.11b
NEARFIELD WAVEFUNCTION
IMAGINARY PART

2.7 Gaussian Envelope Pulse and its Dislocations

We wish now to drive the piston with a pulse which is a small departure from a monochromatic signal, i.e. a quasi-monochromatic pulse. A suitable model is a Gaussian envelope containing n cycles of the carrier within its standard deviation i.e. we take

$$F(t) = e^{-\frac{t^2}{2}\left(\frac{\omega}{\pi n}\right)^2} e^{i\omega t}$$

and compute

$$\Psi(R,Z,T) = \Theta(1-R)\, e^{\frac{-(T-KZ)^2}{2(\pi n)^2}} e^{i(T-KZ)} + \frac{1}{\pi}\int_0^{\pi} d\phi \left(\frac{R\cos\phi - 1}{1+R^2-2R\cos\phi}\right) e^{\frac{-(T-K\sqrt{Z^2+1+R^2-2R\cos\phi})^2}{2(\pi n)^2}} e^{i(T-K\sqrt{Z^2+1+R^2-2R\cos\phi})} \qquad (2.21)$$

On the axis this simplifies to

$$\Psi(0,Z,T) = e^{\frac{-(T-KZ)^2}{2(\pi n)^2}} e^{i(T-KZ)} - e^{\frac{-(T-K\sqrt{Z^2+1})^2}{2(\pi n)^2}} e^{i(T-K\sqrt{Z^2+1})} \qquad (2.22)$$

In the far field

$$\Psi_F(\Theta,\phi) = \frac{2}{\pi K \sin\Theta}\int_0^{\pi} d\phi \cos\phi\, e^{\frac{-(\tau+K\sin\Theta\cos\phi)^2}{2(\pi n)^2}} e^{i(\tau+K\sin\Theta\cos\phi)} \qquad (2.23)$$

for $\theta \neq 0$, and using

$$F'(t) = \omega\left(i - \frac{\omega t}{(\pi n)^2}\right) e^{\frac{-(\omega t)^2}{2(\pi n)^2}} e^{i\omega t}$$

we have

$$\Psi_F(0,\tau) = e^{\frac{-\tau^2}{2(\pi n)^2}}\left(ie^{i\tau} - \frac{\tau}{(\pi n)^2} e^{i\tau}\right) \qquad (2.24)$$

Computations were made with the same centre frequency as was used in the CW case i.e. K = 10.0, and n = 3 (again to model the experimental work). We consider first the far field form. The amplitude at $\theta = 0$, as shown in fig.2.12, is virtually identical to the driving envelope

$$e^{\frac{-(\omega t)^2}{2(\pi n)^2}}$$

apart from a slight broadening (which is just discernable by eye), as expected from eqn.(2.24). The amplitude is generally as one would expect, showing variation in $(\omega t - kr)$ due to the envelope, and variation

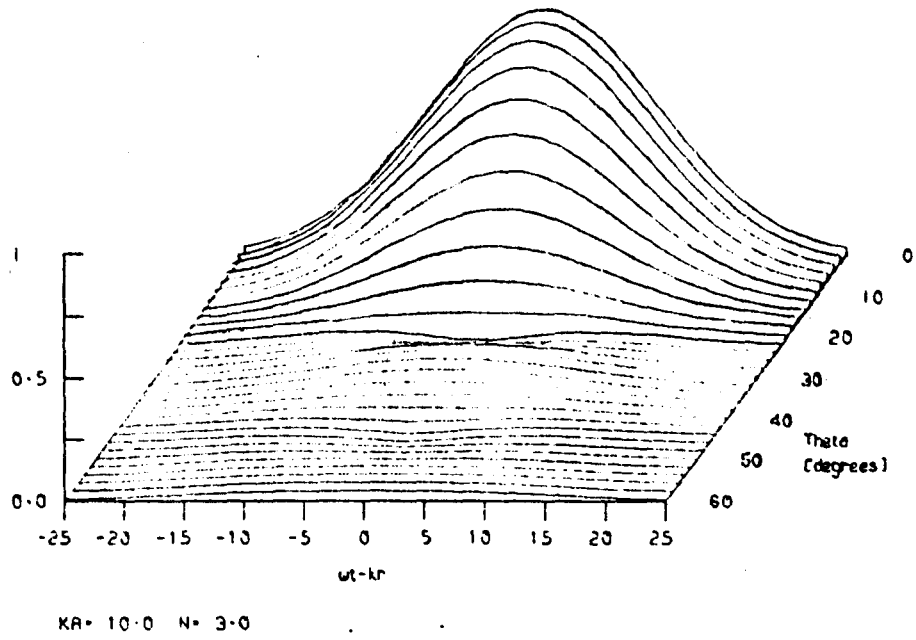


Figure 2.12
REDUCED FARFIELD WAVEFN
AMPLITUDE

in θ similar to that for CW as shown in fig,2.8a, except that now the zeros of amplitude only occur for (ωt - kr) = 0, instead of for all time as in the CW case. The zeros still occur at the CW null angles as near as one can tell. The localization of the zeros has important consequences for the equiphase lines, changing the degenerate comb singularities in fig.2.8c into the two point singularities in fig.2.13b having exactly the canonical pattern shown in fig.10 of NB74 and fig.1.5 of chapter 1.

The effect of the phase singularities on the actual wavefunction is shown in figs.2.14a & b. Consider the imaginary part of the complex wavefunction, as shown in fig.2.14b, near $\theta = 60^{\circ}$. Between (ωt - kr) = ±10 we have 3 crests and 2 troughs. Near $\theta = 45^{\circ}$ the central crest comes to an end and the troughs on either side coalesce into one central trough. Thus we have lost one crest and one trough in a symmetrical fashion, and the remaining crests and troughs have moved in to take up the vacated space. Near $\theta = 23^{\circ}$ the process repeats itself with crests replaced by troughs: the central trough comes to an end and the two crests on either side coalesce into one central crest, so that we lose another crest and trough symmetrically.

The real part of the wavefunction, fig.2.14a, shows exactly the same behaviour, but here the crest and trough pairs disappear anti-symmetrically. One can associate one wavefront with one crest/trough pair (e.g. one could call the peaks of the crests the wavefronts) and we see that near the phase singularities wavefronts come to an end, which is why they are called wavefront dislocations. This occurs along circular loops about the axis such that there is one more wavefront passing outside the loop than there is passing inside it.

Dislocations produced by this particular system must be circles centred on the Z axis and in a plane perpendicular to it. The symmetry constrains the dislocation lines to lie in the wavefront surfaces, hence only pure edge dislocations can occur. This is equivalent to the fact that the symmetry makes the problem essentially two-dimensional (Z and ℛ), and only pure edge dislocations can occur in two dimensions.

We have seen that in the farfield the dislocations are static and occur exactly at the centre of the pulse. Now we consider how

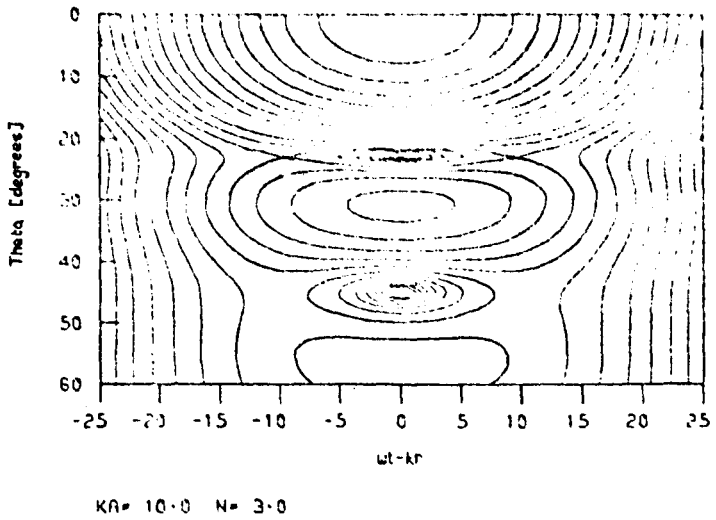


Figure 2.13a
REDUCED FARFIELD WAVEFN
CONTOURS OF LOG[AMP]

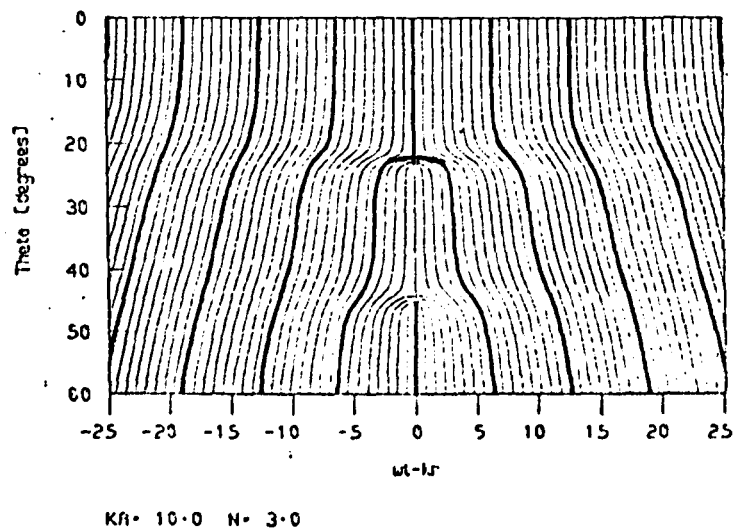


Figure 2.13b
REDUCED FARFIELD WAVEFN
CONTOURS OF PHASE

Compare heavy lines at $\pi/2$ with figure 2.8c

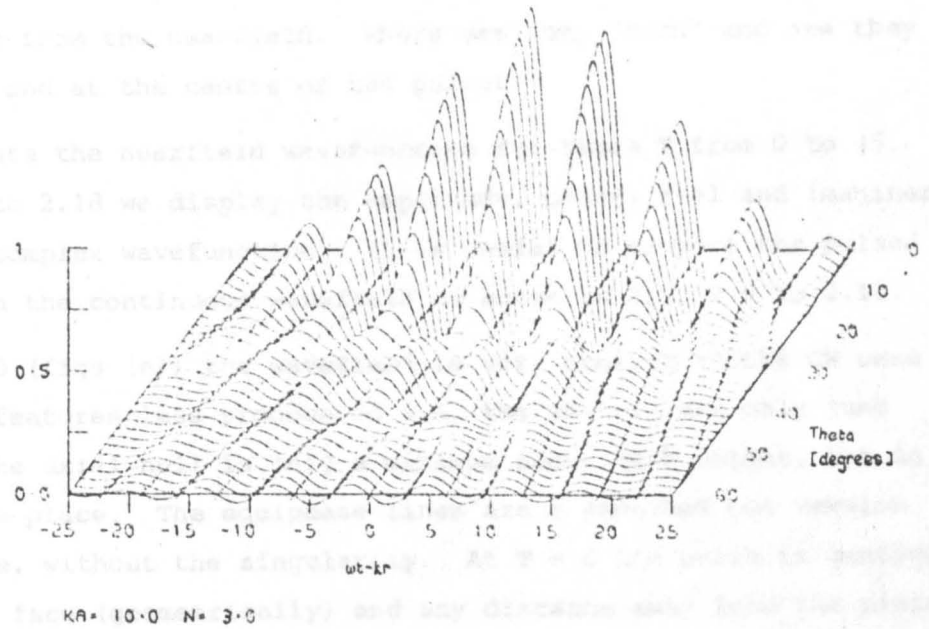


Figure 2.14a
REDUCED FARFIELD WAVEFN
REAL PART

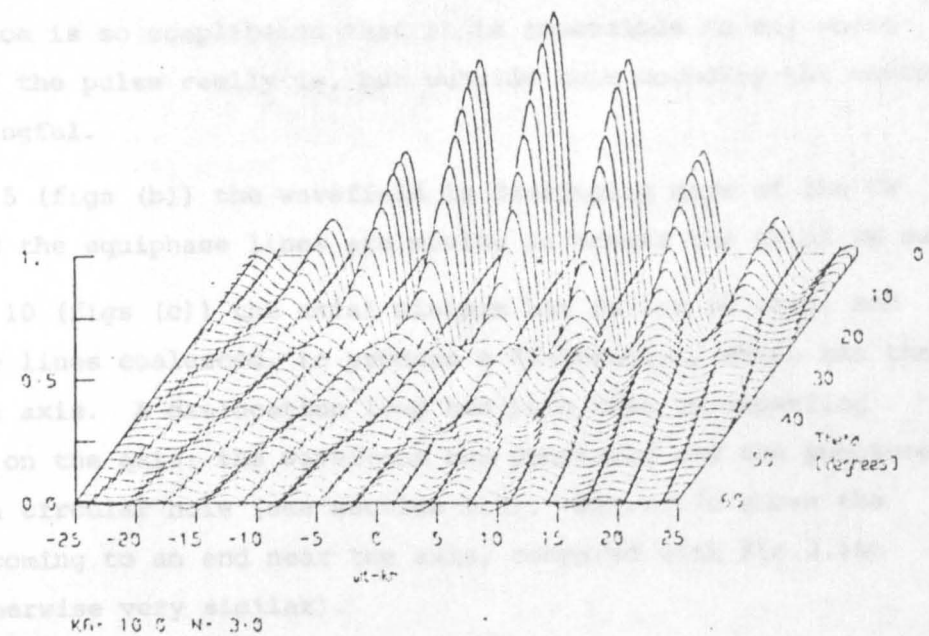


Figure 2.14b
REDUCED FARFIELD WAVEFN
IMAGINARY PART

they got there from the nearfield. Where are they "born" and are they always static and at the centre of the pulse?

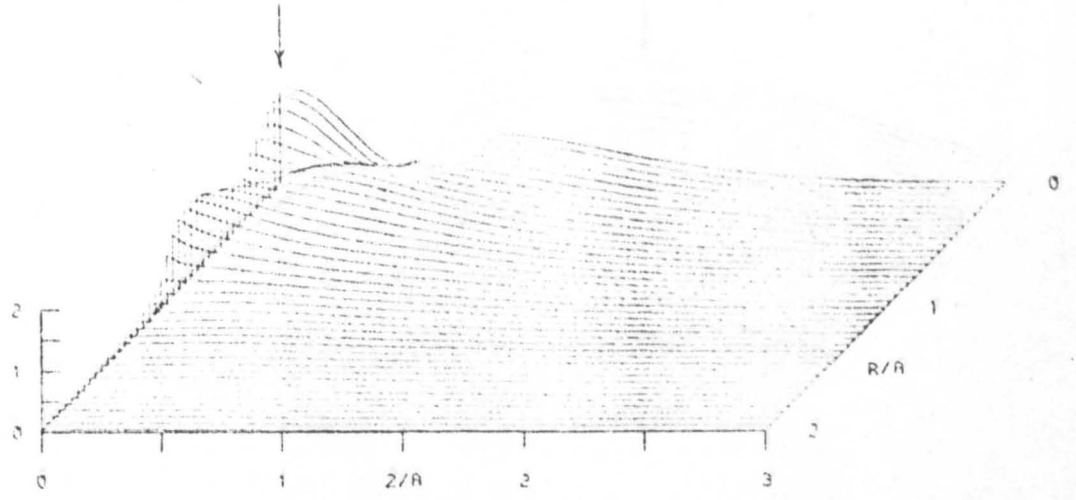
We compute the nearfield wavefunction for times T from 0 to 45. In figs.2.15 to 2.18 we display the amplitude, phase, real and imaginary parts of the complex wavefunction. It is useful to compare the pulsed wavefield with the continuous wavefield as shown in figs.2.9 to 2.11.

At T = 0 (figs (a)) the wavefield is very similar to the CW case but with the features less pronounced i.e. the valleys are only just visible and the axial null is only a minimum above zero height, but in about the same place. The equiphase lines are a smoothed out version of the CW case, without the singularity. At T = 0 the pulse is centred on the piston face (geometrically) and any distance away from the piston the amplitude is only just beginning to build up. The smoothness of the wavefield suggests that the interesting diffraction phenomena in the CW wavefield actually take time to build up. We shall see that as time progresses the wavefronts become twisted and tangled up as more and more "rays" have time to interact. The arrow indicates where T - KZ = 0 i.e. where one would geometrically expect the centre of the pulse to occur along the axis. Inside the near/far boundary (situated at Z ≈ 1.4) the diffraction is so complicated that it is impossible to say where the centre of the pulse really is, but outside this boundary the concept becomes meaningful.

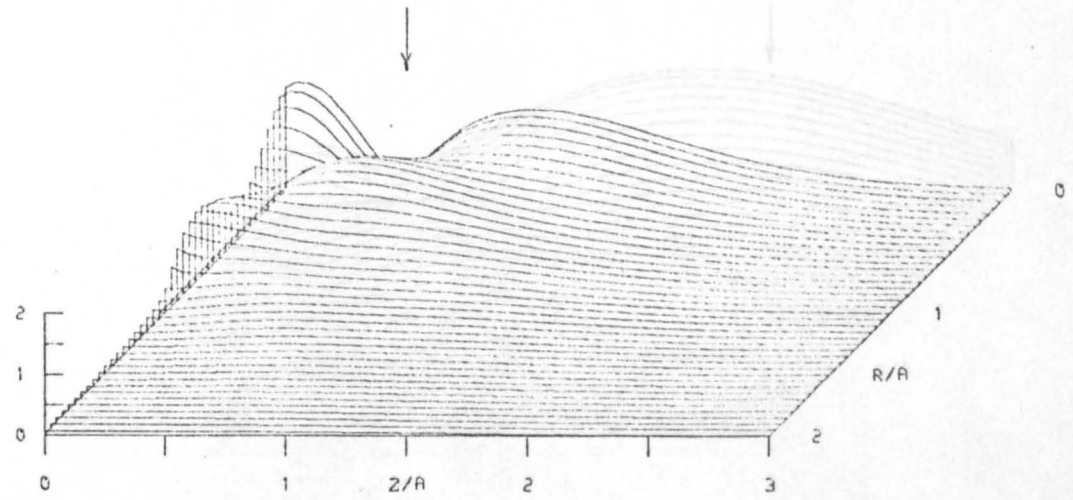
At T = 5 (figs (b)) the wavefield is developing more of the CW structure and the equiphase lines are moving in toward the axial CW null.

At T = 10 (figs (c)) the axial minimum has fallen to zero, and the equiphase lines coalesced, to produce a dislocation, which has then moved off the axis. A dislocation loop has been born by expanding from a point on the axis: the wavefront has punctured and the puncture opened into a circular hole (see section 1.7). Fig.2.17c shows the first crest coming to an end near the axis, compared with fig.2.18b (which is otherwise very similar).

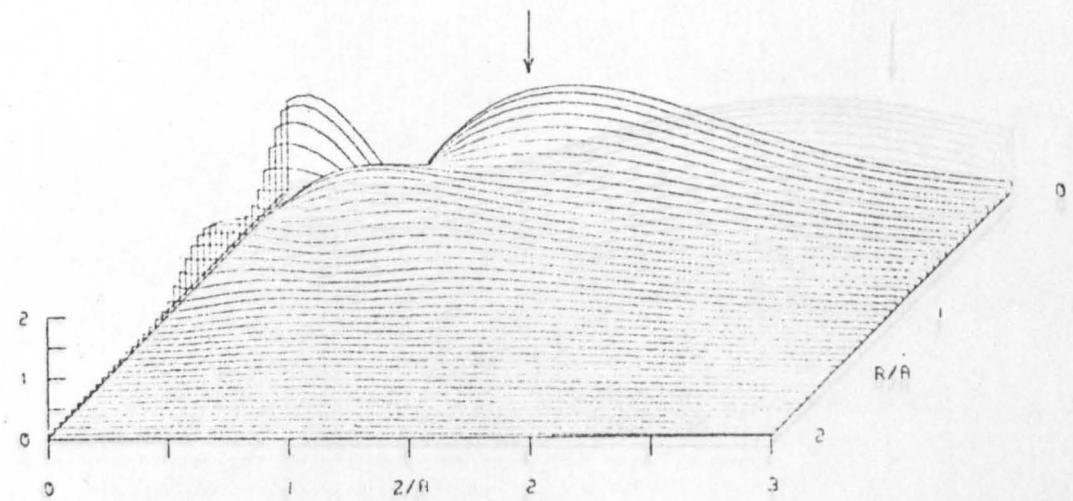
At T = 15 (figs (d)) the zero of amplitude just off the axis is more clearly visible in fig.2.15d, along with a new dip just in front of the piston. We see from the phase map, fig.2.16d, that the dislocation near the axis has moved slightly further away from the axis (and the equiphase lines have rotated around it), and a new pair of dislocations,



(a) TIME= 0.0 KA= 10.0 N= 3.0

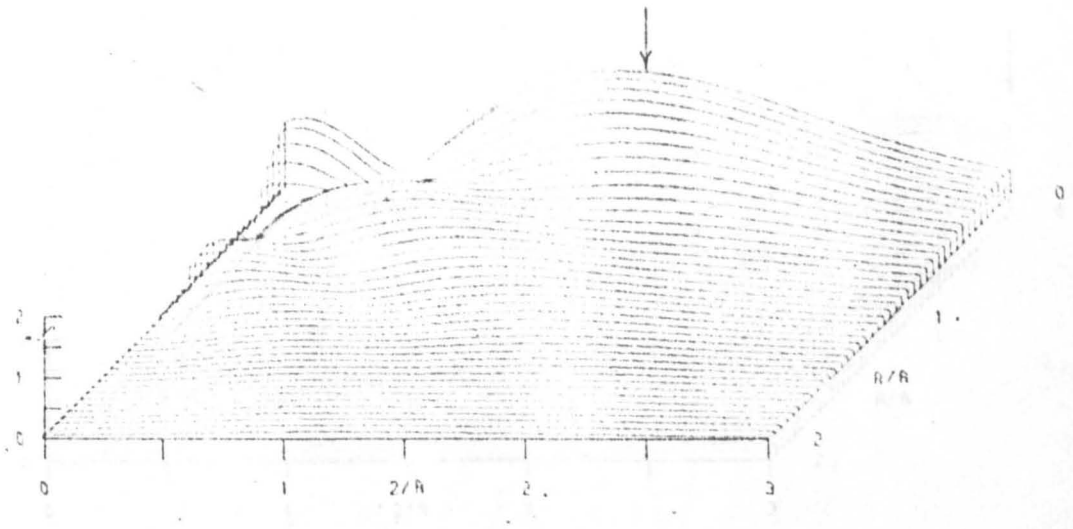


(b) TIME= 5.0 KA= 10.0 N= 3.0

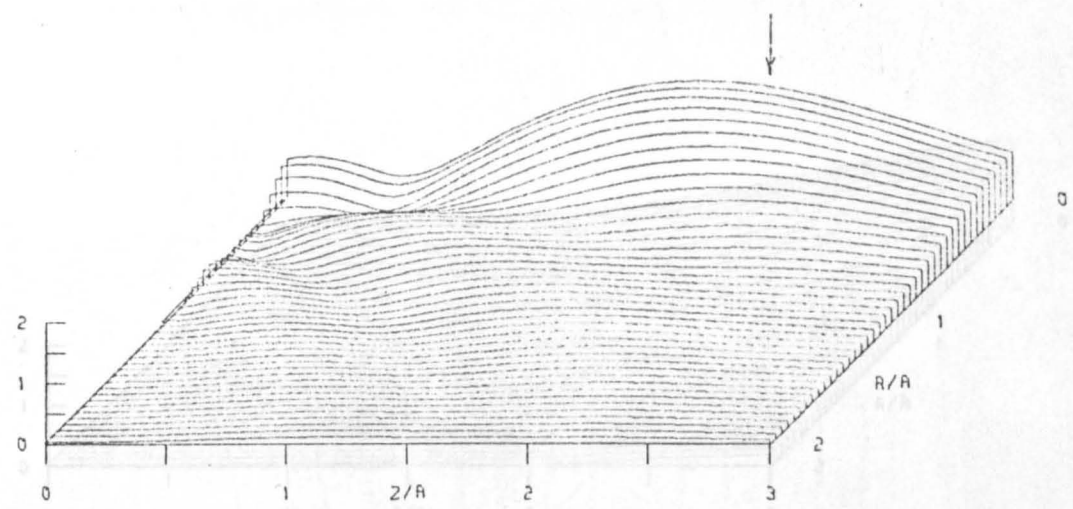


(c) TIME= 10.0 KA= 10.0 N= 3.0

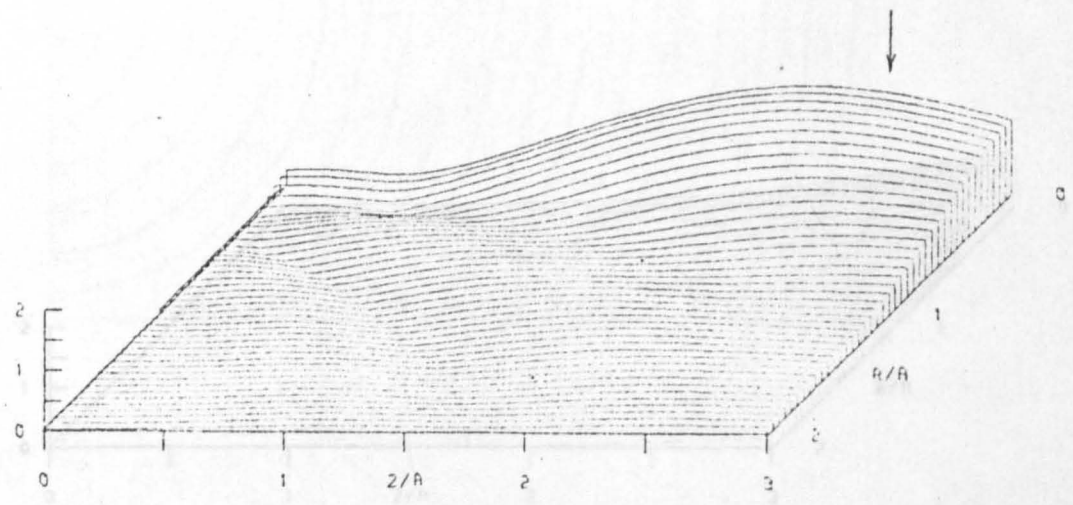
Figure 2.15
NEARFIELD WAVEFUNCTION
AMPLITUDE



(d) TIME= 15.0 KA= 10.0 N= 3.0

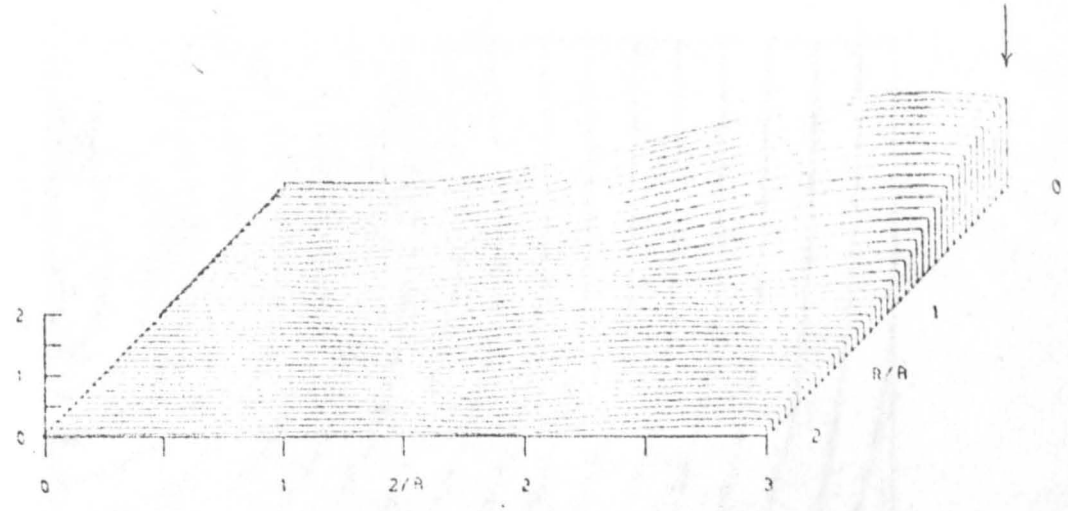


(e) TIME= 20.0 KA= 10.0 N= 3.0

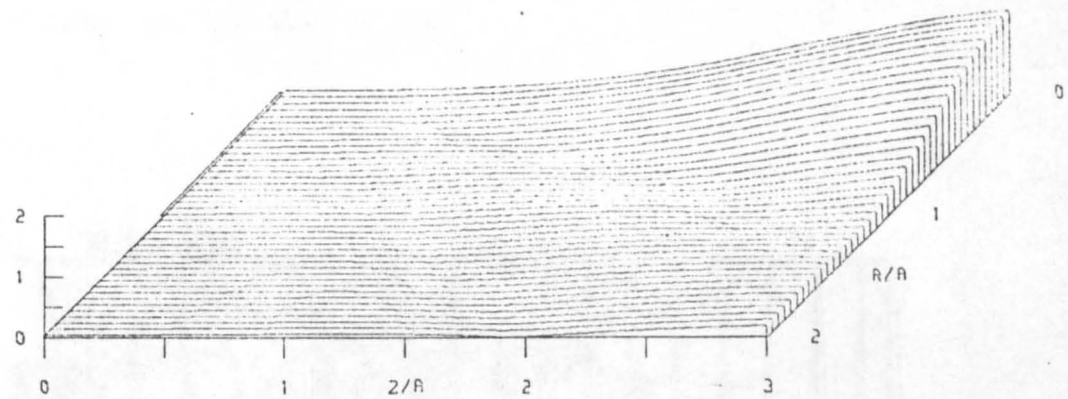


(f) TIME= 25.0 KA= 10.0 N= 3.0

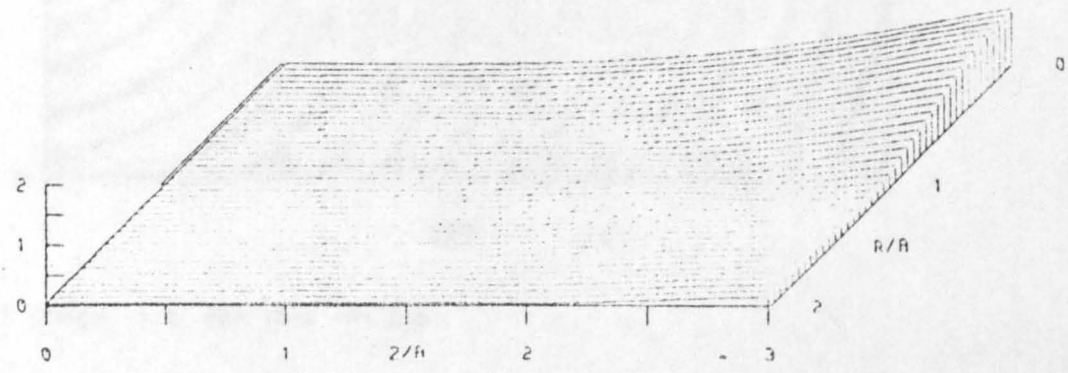
Figure 2.15
NEARFIELD WAVEFUNCTION
AMPLITUDE



(g) TIME= 30.0 KA= 10.0 N= 3.0

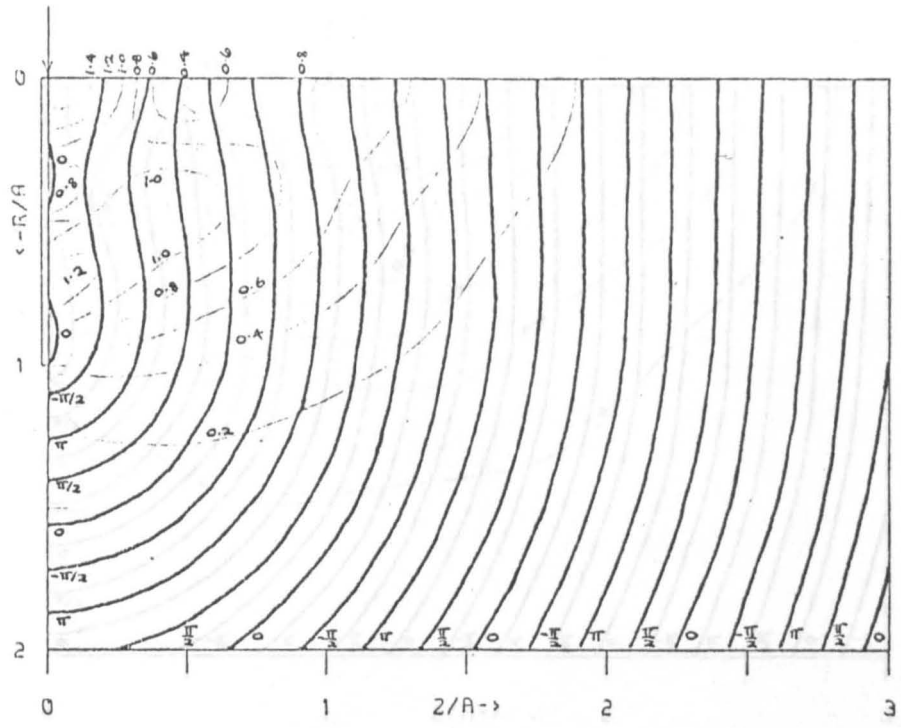


(h) TIME= 35.0 KA= 10.0 N= 3.0

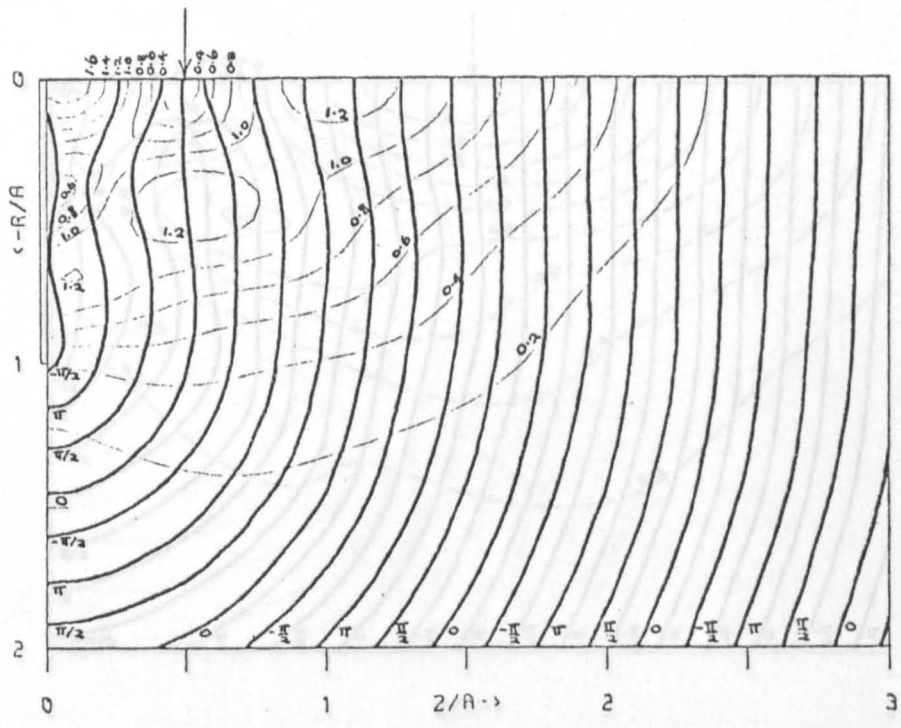


(i) TIME= 40.0 KA= 10.0 N= 3.0

Figure 2.15
NEARFIELD WAVEFUNCTION
AMPLITUDE

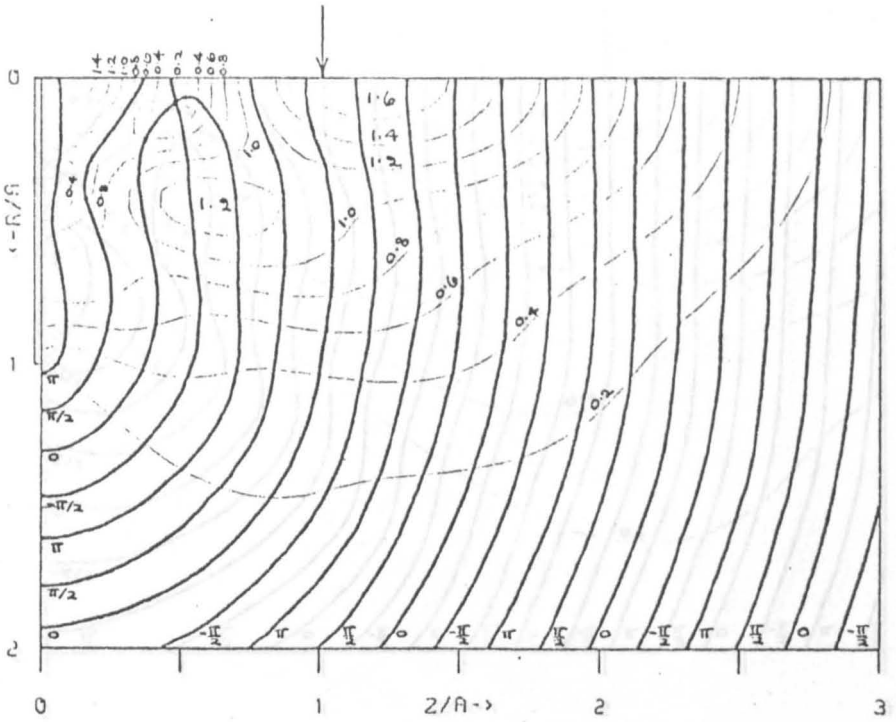


(a) TIME= 0·0 KA= 10·0 N= 3·0

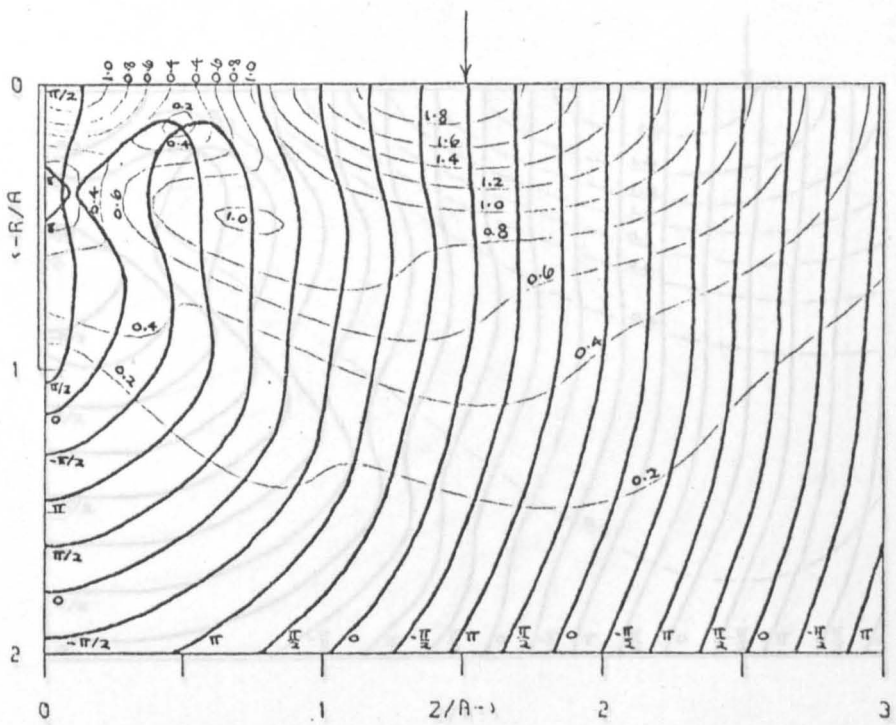


(b) TIME= 5·0 KA= 10·0 N= 3·0

Figure 2.16 Nearfield Wavefunction - Contours
Phase (thick lines labelled by fractions of Pi)
Amplitude (thin lines labelled by decimals)

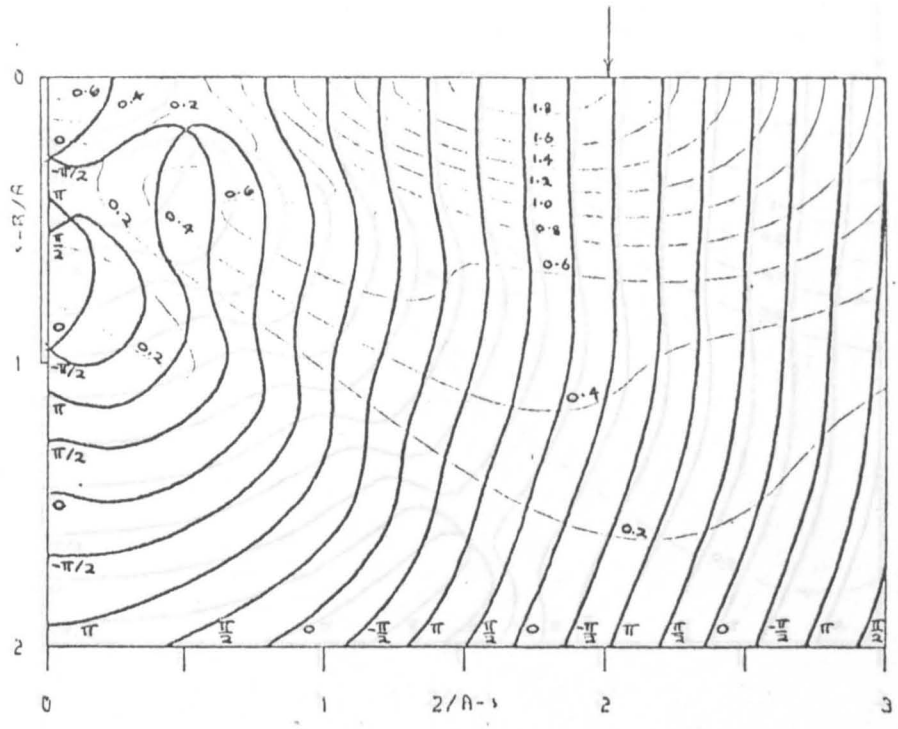


(c) TIME= 10·0 KA= 10·0 N= 3·0

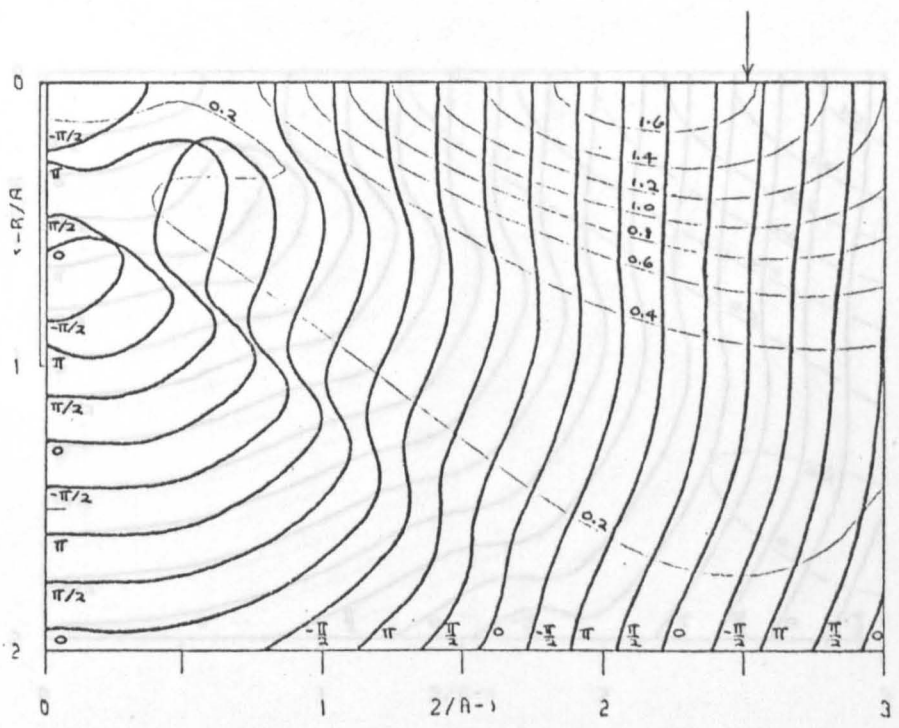


(d) TIME= 15·0 KA= 10·0 N= 3·0

Figure 2.16 Nearfield Wavefunction - Contours
Phase (thick lines labelled by fractions of Pi)
Amplitude (thin lines labelled by decimals)

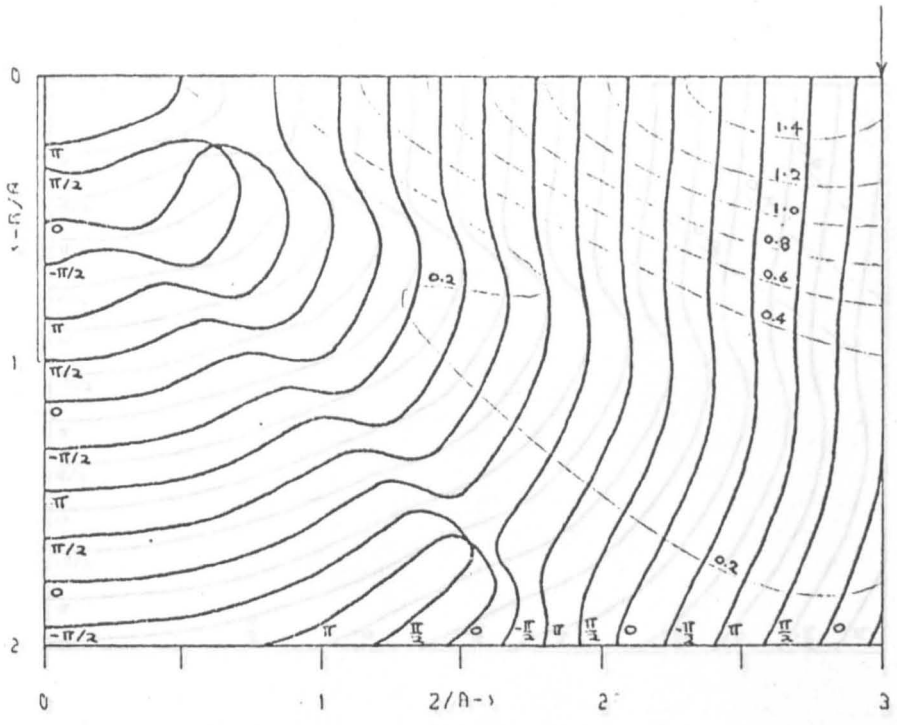


(e) TIME= 20.0 KA= 10.0 N= 3.0

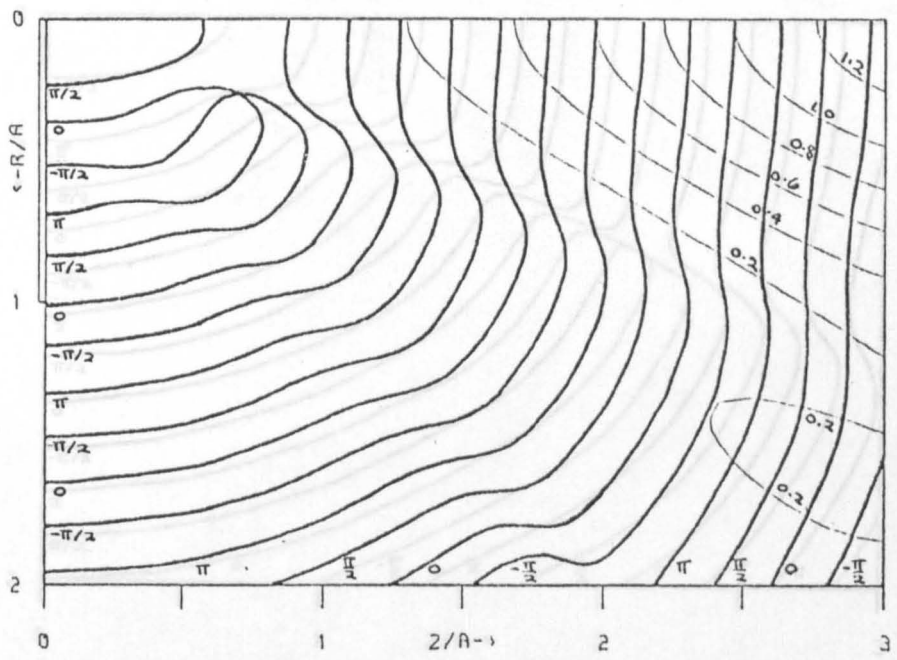


(f) TIME= 25.0 KA= 10.0 N= 3.0

Figure 2.16 Nearfield Wavefunction - Contours
Phase (thick lines labelled by fractions of Pi)
Amplitude (thin lines labelled by decimals)

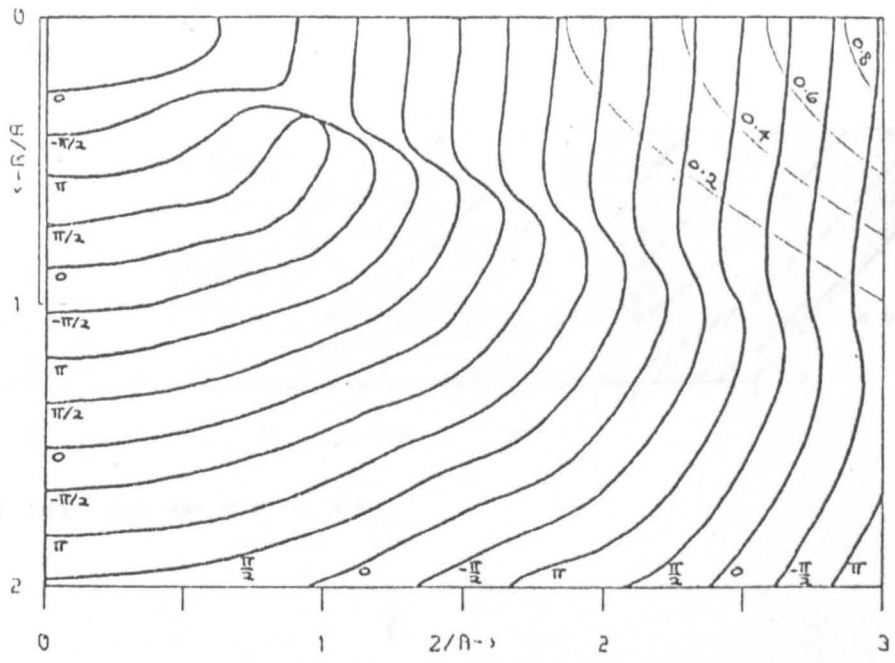


(g) TIME= 30.0 KA= 10.0 N= 3.0

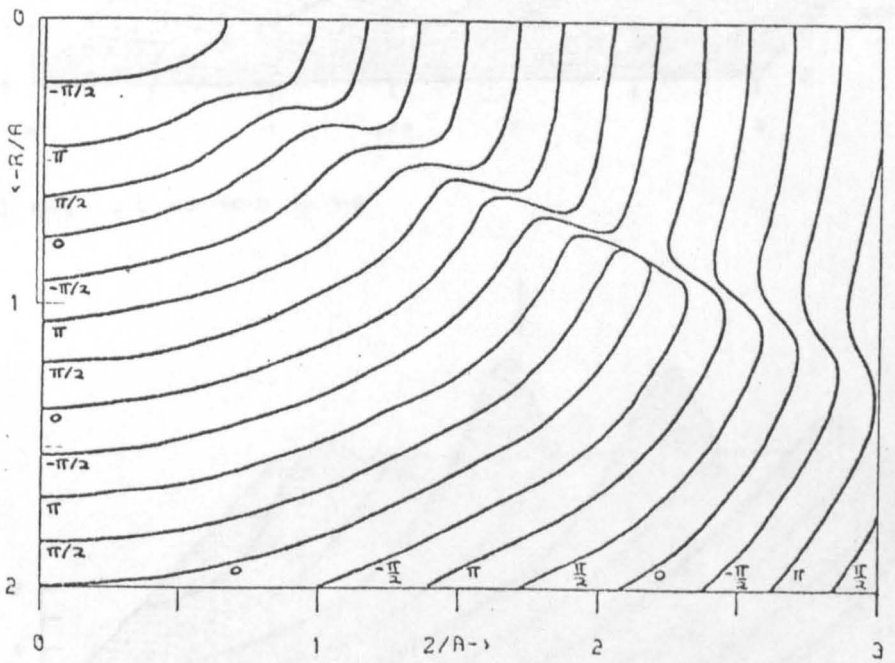


(h) TIME= 35.0 KA= 10.0 N= 3.0

Figure 2.16 Nearfield Wavefunction - Contours
Phase (thick lines labelled by fractions of Pi)
Amplitude (thin lines labelled by decimals)

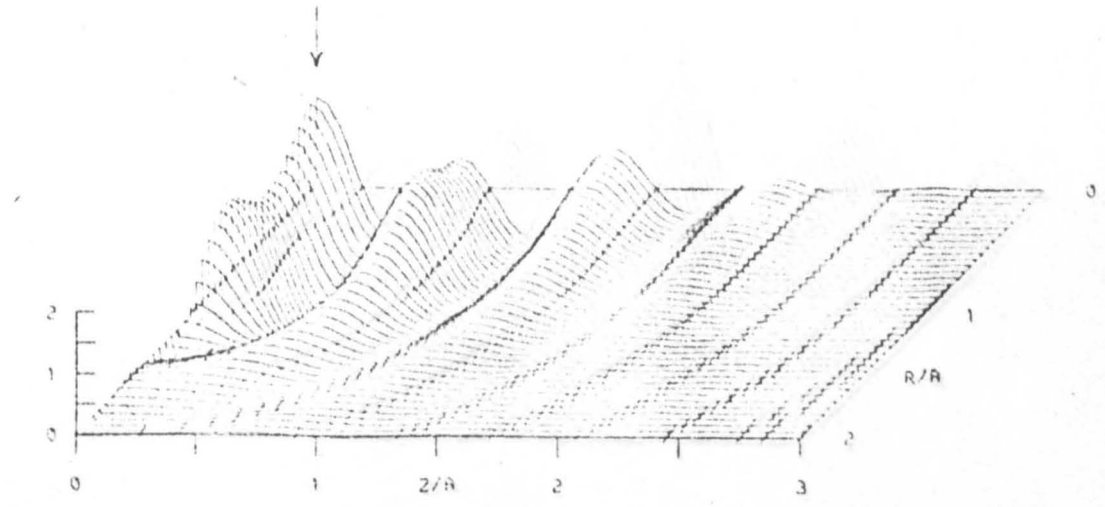


(i) TIME= 40·0 KA= 10·0 N= 3·0

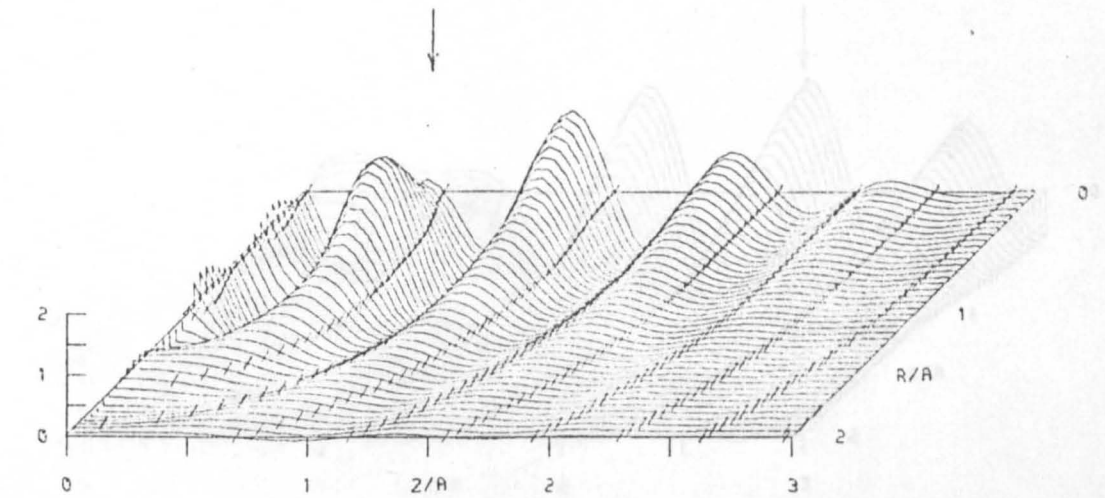


(j) TIME= 45·0 KA= 10·0 N= 3·0

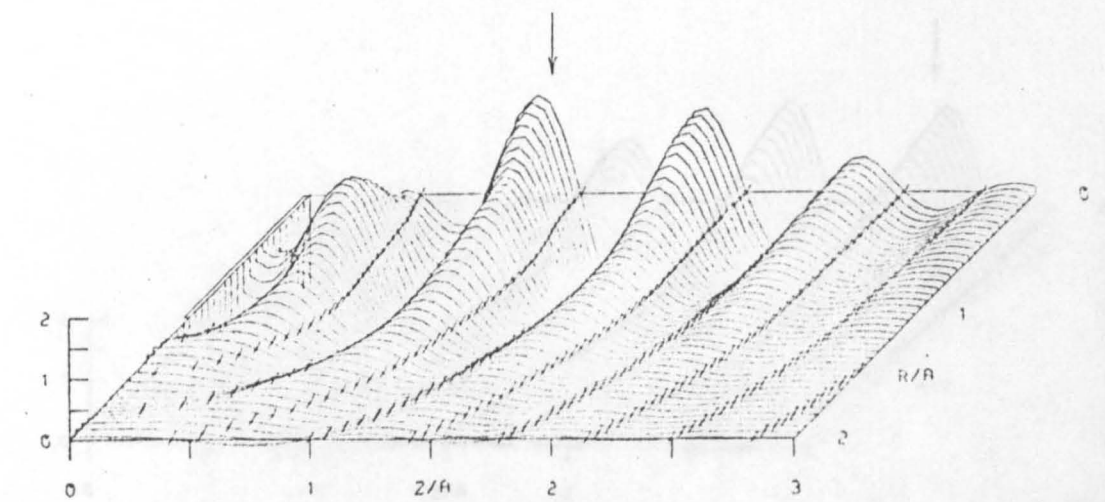
Figure 2.16 Nearfield Wavefunction - Contours
Phase (thick lines labelled by fractions of Pi)
Amplitude (thin lines labelled by decimals)



(a) TIME= 0.0 KA= 10.0 N= 3.0

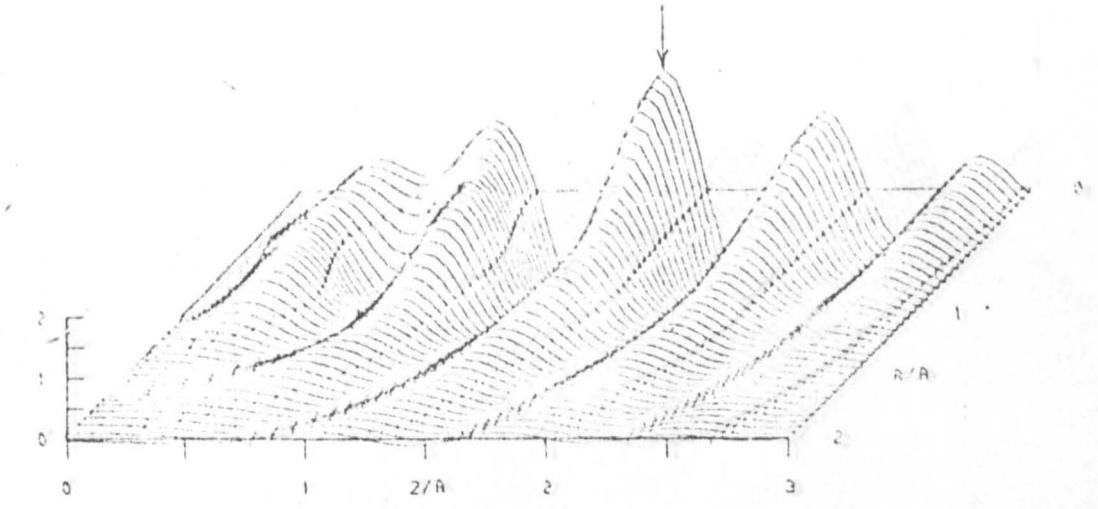


(b) TIME= 5.0 KA= 10.0 N= 3.0

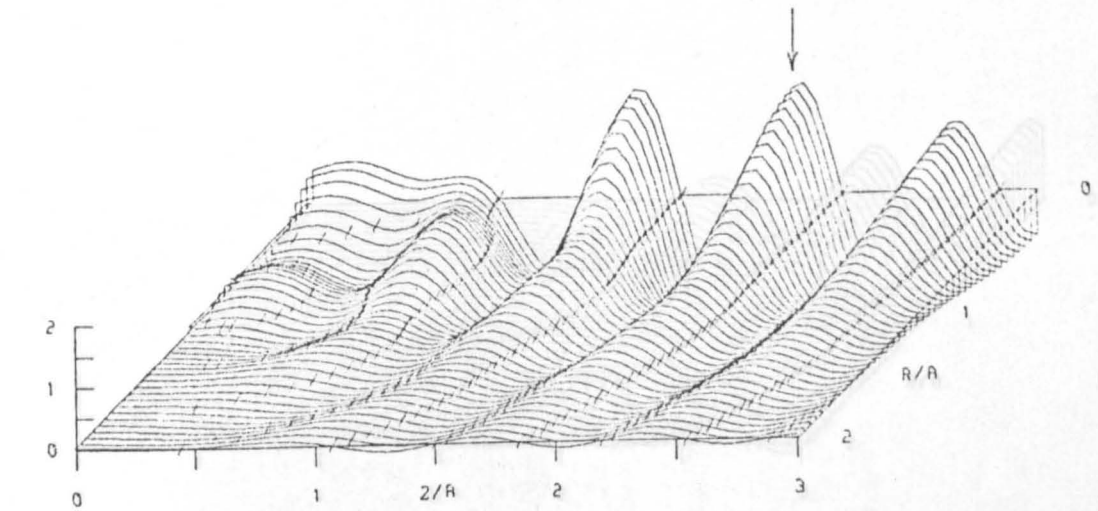


(c) TIME= 10.0 KA= 10.0 N= 3.0

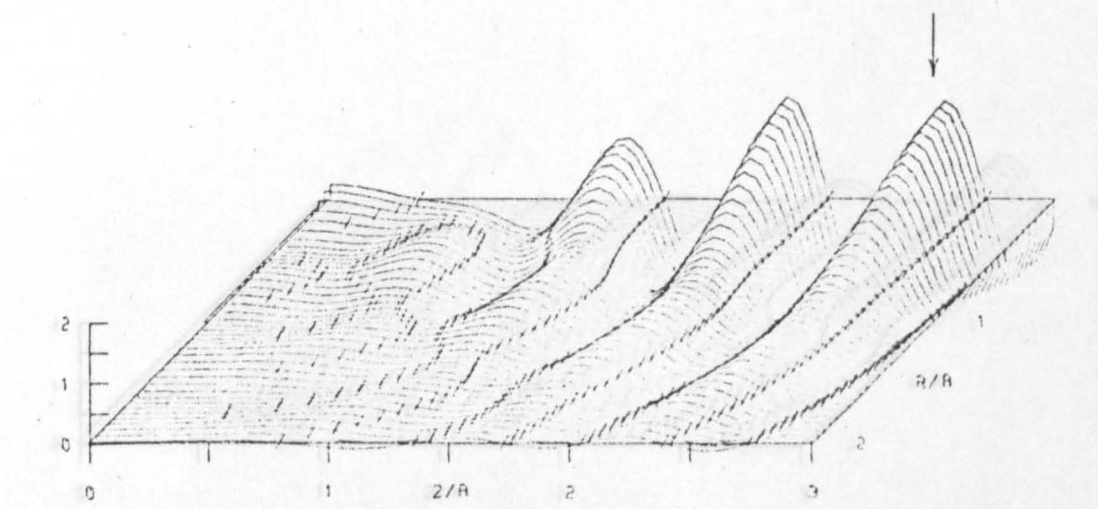
Figure 2.17
NEARFIELD WAVEFUNCTION
REAL PART



(d) TIME= 15.0 KA= 10.0 N= 3.0

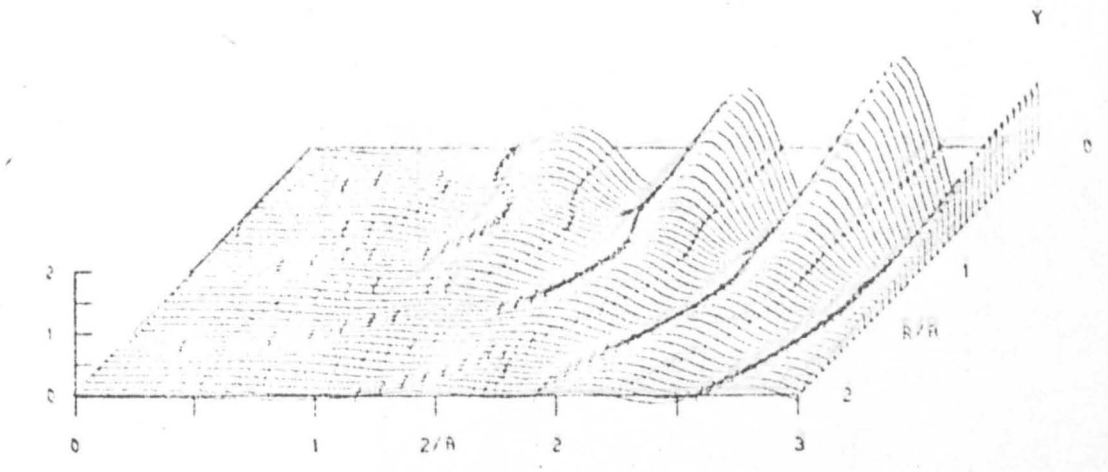


(e) TIME= 20.0 KA= 10.0 N= 3.0

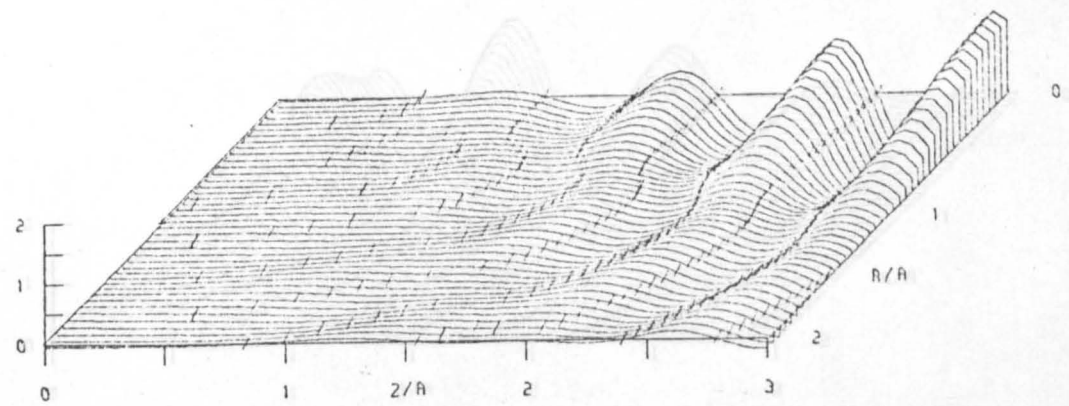


(f) TIME= 25.0 KA= 10.0 N= 3.0

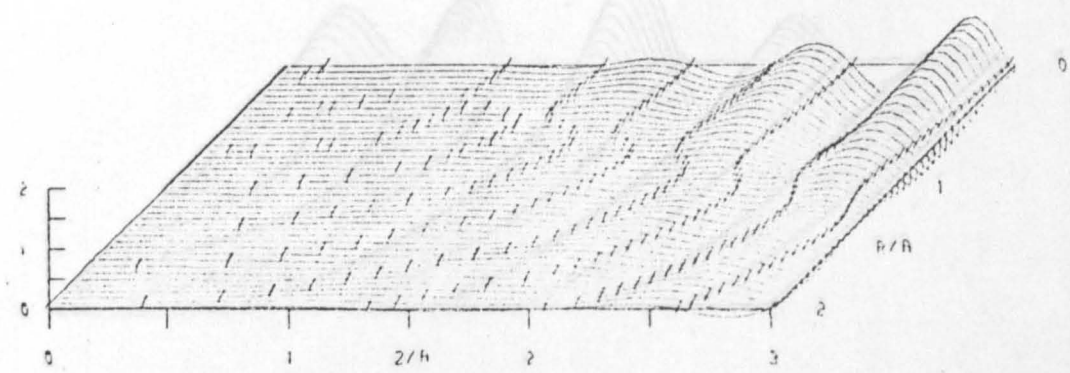
Figure 2.17
NEARFIELD WAVEFUNCTION
REAL PART



(g) TIME= 30.0 KA= 10.0 N= 3.0

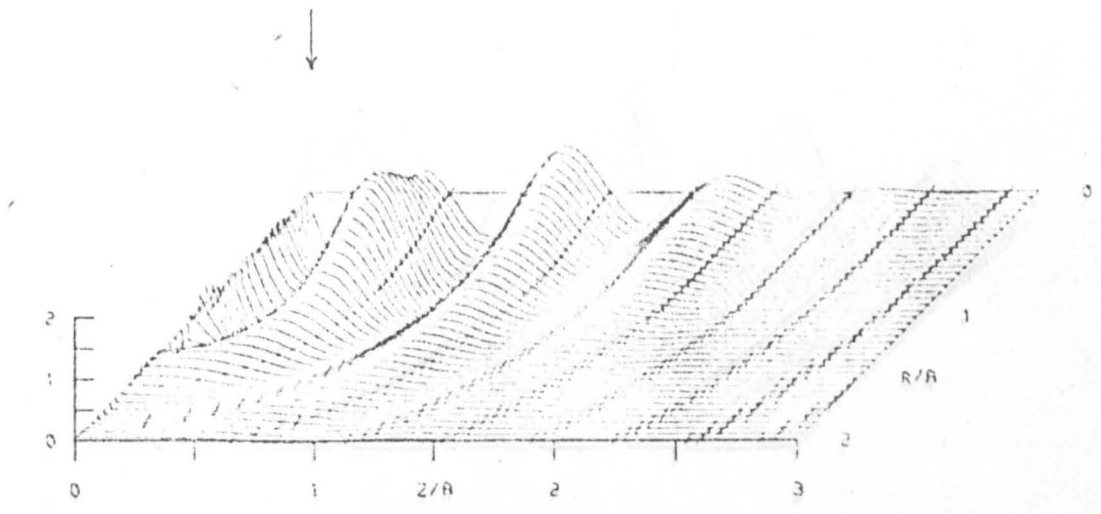


(h) TIME= 35.0 KA= 10.0 N= 3.0

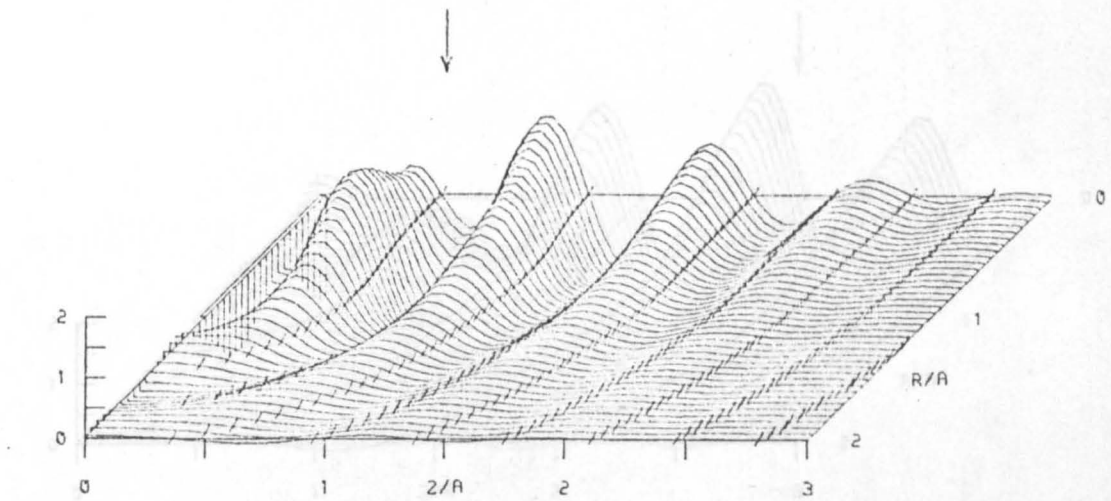


(i) TIME= 40.0 KA= 10.0 N= 3.0

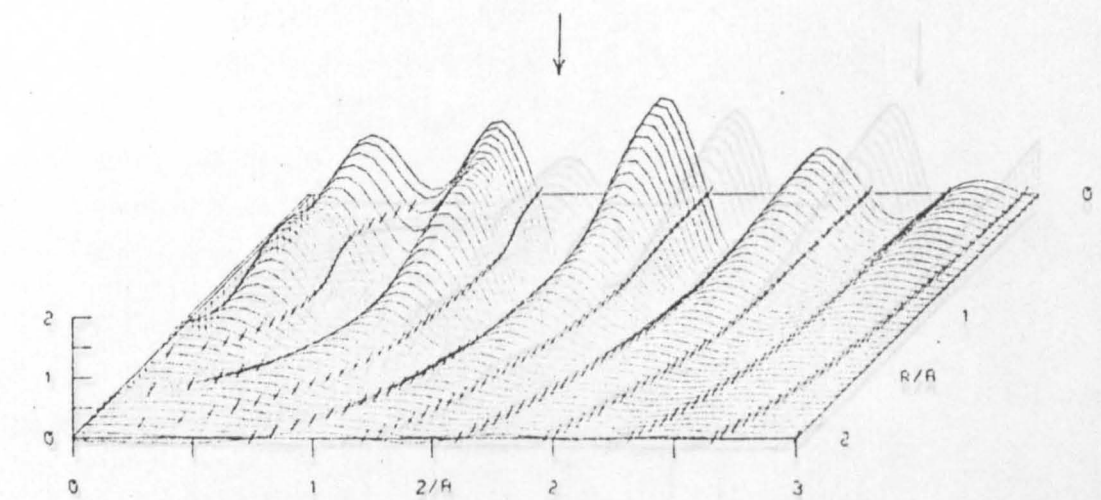
Figure 2.17
NEARFIELD WAVEFUNCTION
REAL PART



(a) TIME= 0.0 KA= 10.0 N= 3.0

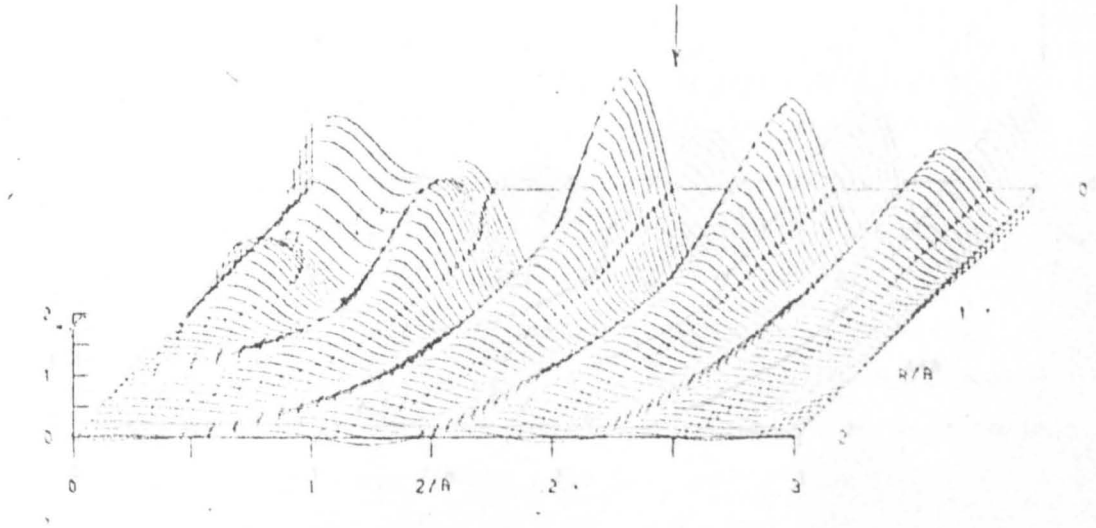


(b) TIME= 5.0 KA= 10.0 N= 3.0

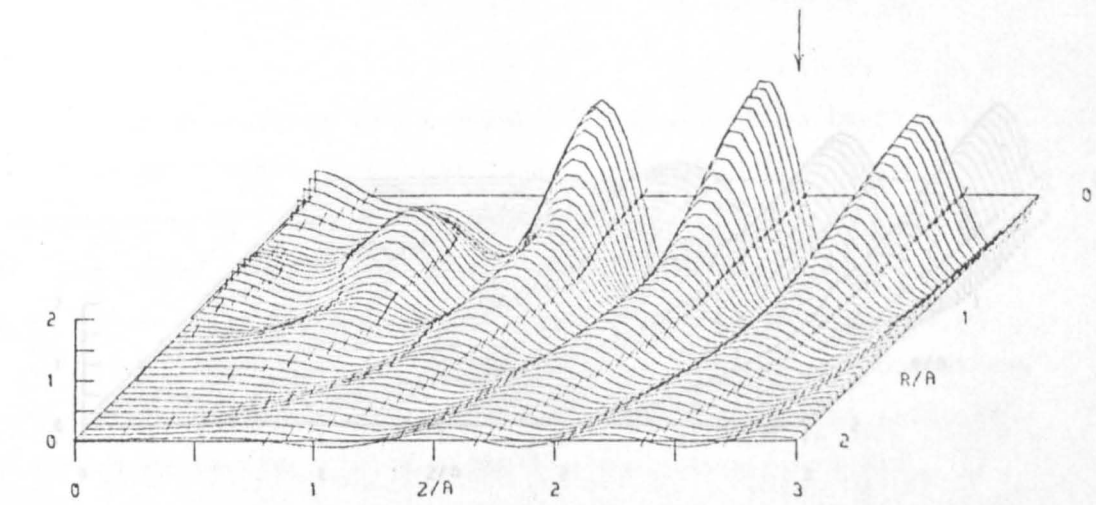


(c) TIME= 10.0 KA= 10.0 N= 3.0

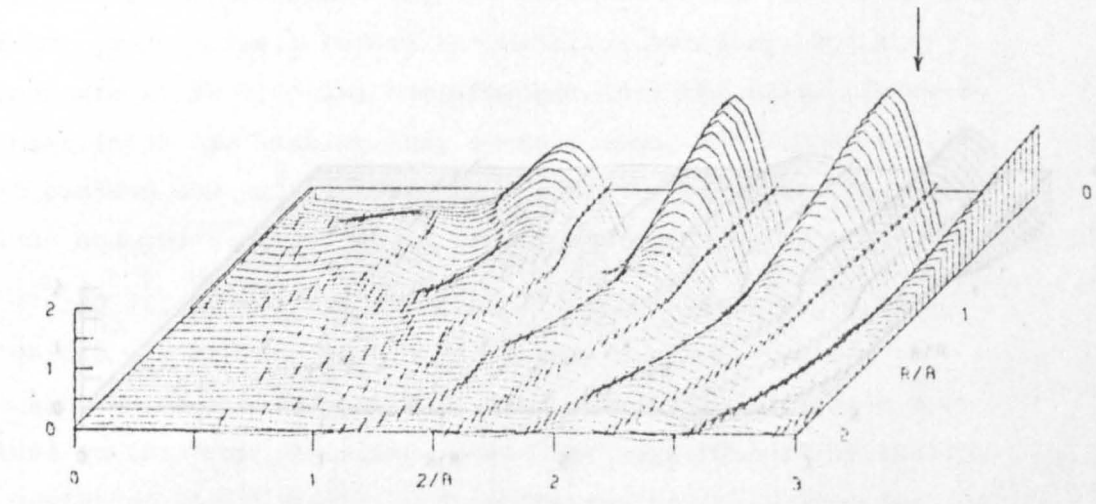
Figure 2.18
NEARFIELD WAVEFUNCTION
IMAGINARY PART



(d) TIME= 15.0 KA= 10.0 N= 3.0

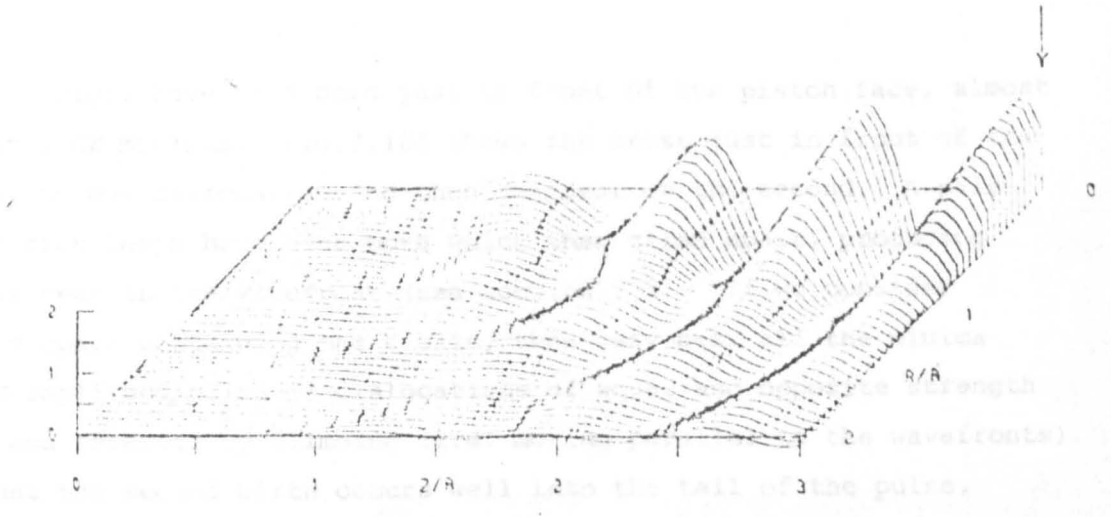


(e) TIME= 20.0 KA= 10.0 N= 3.0

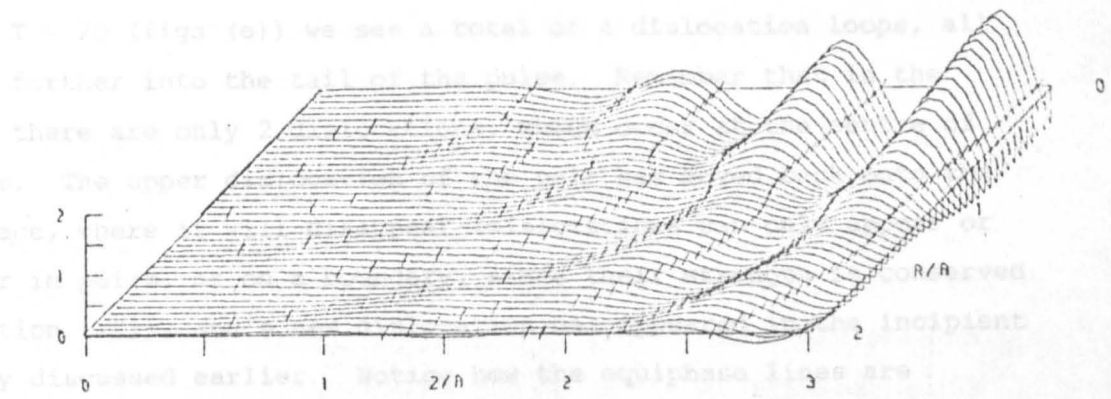


(f) TIME= 25.0 KA= 10.0 N= 3.0

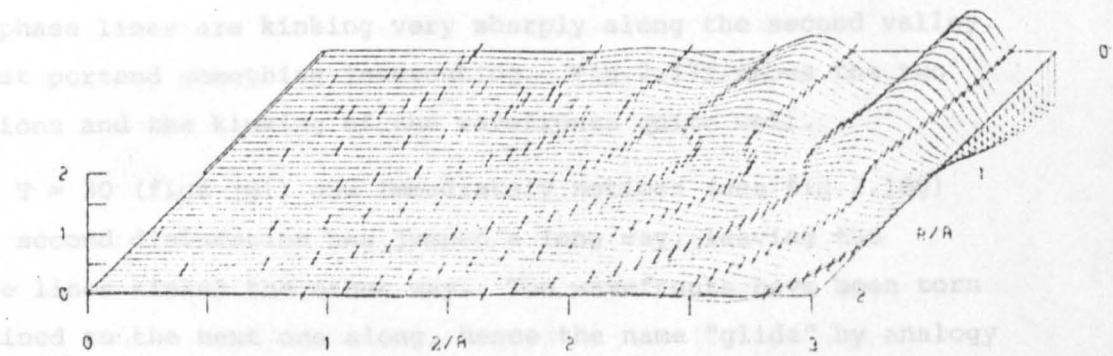
Figure 2.18
NEARFIELD WAVEFUNCTION
IMAGINARY PART



(g) TIME 30 0 KA= 10 0 N= 3·0



(h) TIME= 35·0 KA= 10·0 N= 3·0



(i) TIME= 40 0 KA= 10 0 N= 3·0

Figure 2.18
NEARFIELD WAVEFUNCTION
IMAGINARY PART

of opposite sign, have been born just in front of the piston face, almost exactly at a CW minimum. Fig.2.18d shows the crest just in front of the piston end on one dislocation and then reappear on the second. A pair of dislocation loops have been born which then climb apart, producing an annular tear in the wavefront (see section 1.7). If we consider a plane in space containing the Z axis, then very near all the minima of the CW amplitude, pairs of dislocations of equal and opposite strength are born and separate by climbing (i.e. moving parallel to the wavefronts). Notice that the second birth occurs well into the tail of the pulse, and the first dislocation has dropped well back into the tail also. The centre of the pulse is slightly behind its geometrical position, and seems to remain so until the farfield, when it does occur at $\omega t - kr = 0$.

At T = 20 (figs (e)) we see a total of 4 dislocation loops, all dropping further into the tail of the pulse. Remember that in the farfield there are only 2 dislocations, which occur at the centre of the pulse. The upper dislocation of the pair has moved back onto the piston face, where it will disappear (dislocations can only appear or disappear in pairs, or on a boundary, since their strength is conserved (see section 1.2)), and a new dislocation has appeared in the incipient CW valley discussed earlier. Notice how the equiphase lines are beginning to kink along the valleys of the amplitude.

By T = 25 (figs (f)) the two dislocations on the piston face have disappeared, but the remaining two have not moved very much. The centre of the pulse is well beyond the near/far boundary, but the dislocations are still dropping back further into the tail. However, the equiphase lines are kinking very sharply along the second valley, which must portend something interesting. Fig.2.17f shows the two dislocations and the kinking of the wavefronts quite well.

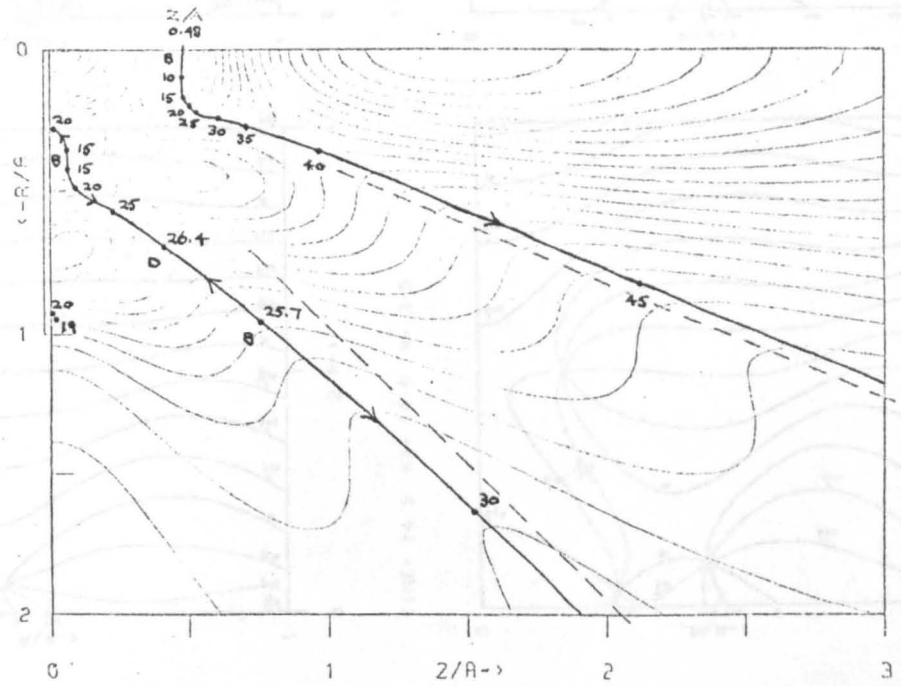
At T = 30 (figs (g)) one immediately notices (see fig.2.16g) that the second dislocation has jumped a long way, leaving the equiphase lines kinked the other way. The wavefronts have been torn and rejoined to the next one along, hence the name "glide" by analogy with the motion of dislocation lines in sheared crystals (see section 1.6). The equiphase lines around the dislocation now take on the canonical pattern as in the far field, but it is still in the tail of the pulse.

At T = 35 (figs (h)) the second dislocation has moved off with the pulse, but even by T = 40 (figs (i)) the first dislocation has not moved much. Figs.2.15(i) & 2.16(i) show just how far into the tail it has got. Finally, at T = 45 (fig 2.16(j) only), with a feeling of desperation since the amplitude is so low that numerical errors are beginning to show up in the phase lines, we find that the first dislocation to be born has finally made an effort to catch up with the pulse!

In fig.2.19 we summarize the behaviour of the dislocations by plotting the trajectories along which they move, and their times, superimposed on the CW amplitude contours. We see that roughly speaking dislocations are born in CW amplitude minima, travel up the valleys to saddles, and then down the valleys on the other side. But only the two valleys leading into the far field carry dislocations, not the valleys and saddle joining the two minima.

We have seen that the phase maps are the most informative about the behaviour of dislocations, therefore we will study in greater detail the phase maps associated with some of the interesting events. Fig.2.20 shows the birth of the dislocation pair near the piston face, and also the incipient dislocation which is seen to move upwards and towards the piston face from somewhere. We shall investigate this incipient dislocation further in the next chapter. Fig.2.21a shows the second dislocation just before its rapid glide. The phase lines kink so that they practically coalesce along a line, until at time T = 26.0 another dislocation seems to have appeared associated with the glide process. Fig.2.21b shows the same dislocation after its rapid glide. The phase lines become progressively less sharply kinked, and the dislocation slowly and smoothly catches up with the pulse, whose "geometrical position" is marked. There is no evidence of any further jumps.

Fig.2.22 shows the rapid glide process in greater detail. Fig.2.22a shows the dislocation which is about to glide, plus two more dislocations spaced along the line of glide. A pair must have been born between T = 25.5 and 26.0, and glided apart. The original dislocation does not move a great deal, but the newly born pair separate rapidly, and one of the pair approaches the original dislocation.



KA= 10·0 N= 3·0

Figure 2.19

NEARFIELD WAVEFUNCTION DISLOCATION TRAJECTORY

showing directions and times, superimposed on CW amplitude contours (B is a birth event, D is a death event)

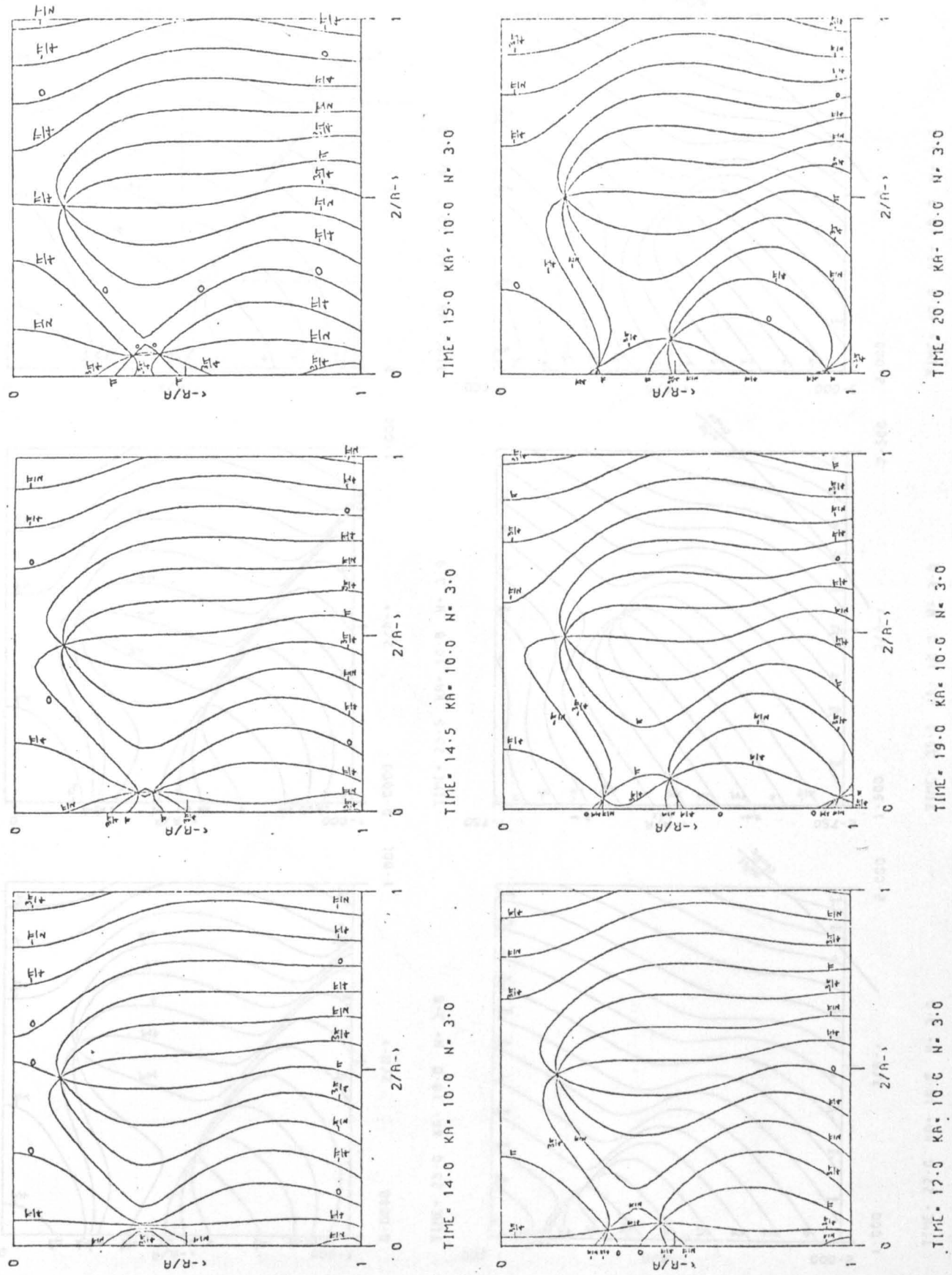


Figure 2.20 Details of Birth near Piston

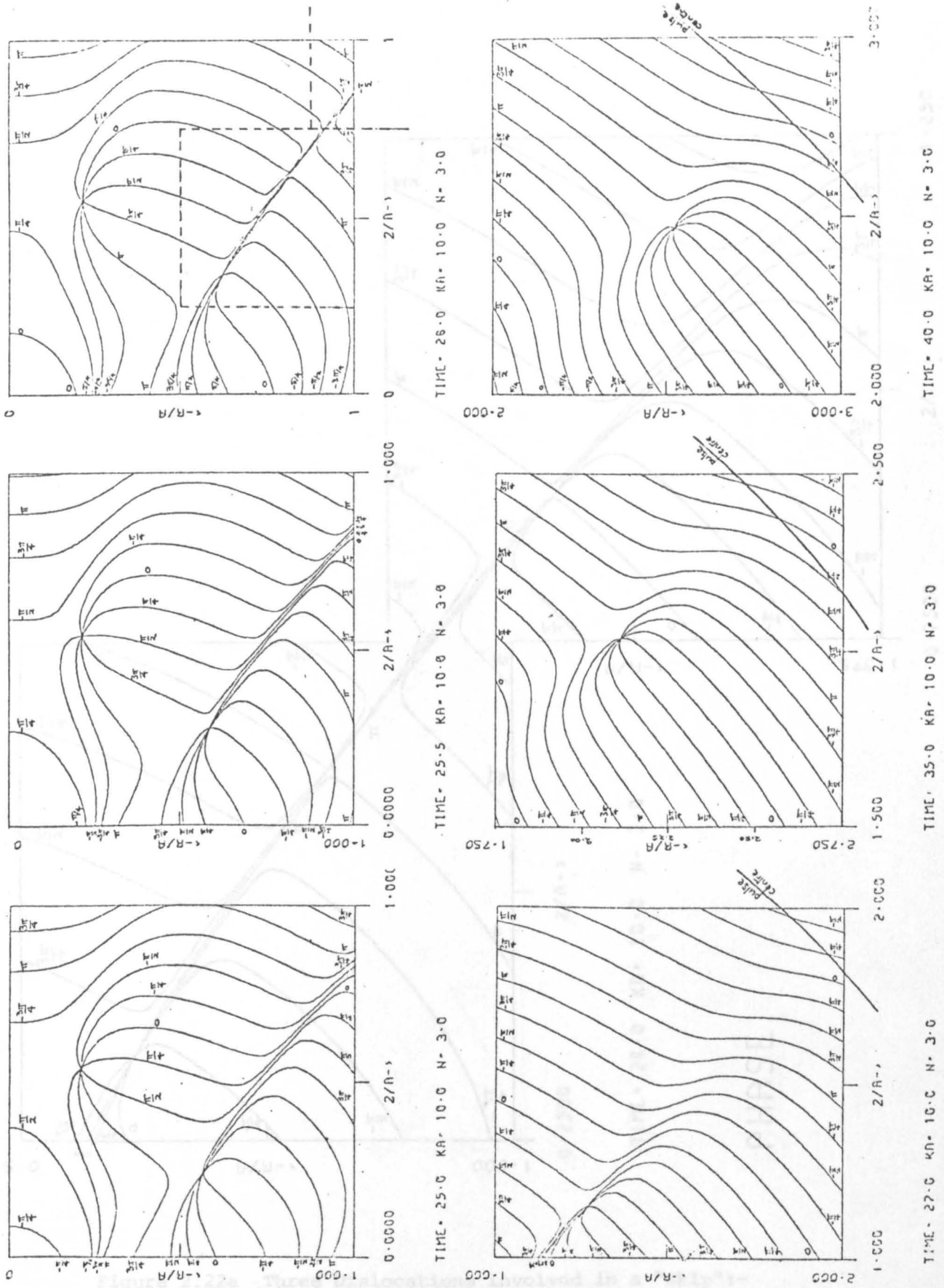


Figure 2.21 Before and After the Rapid Glide

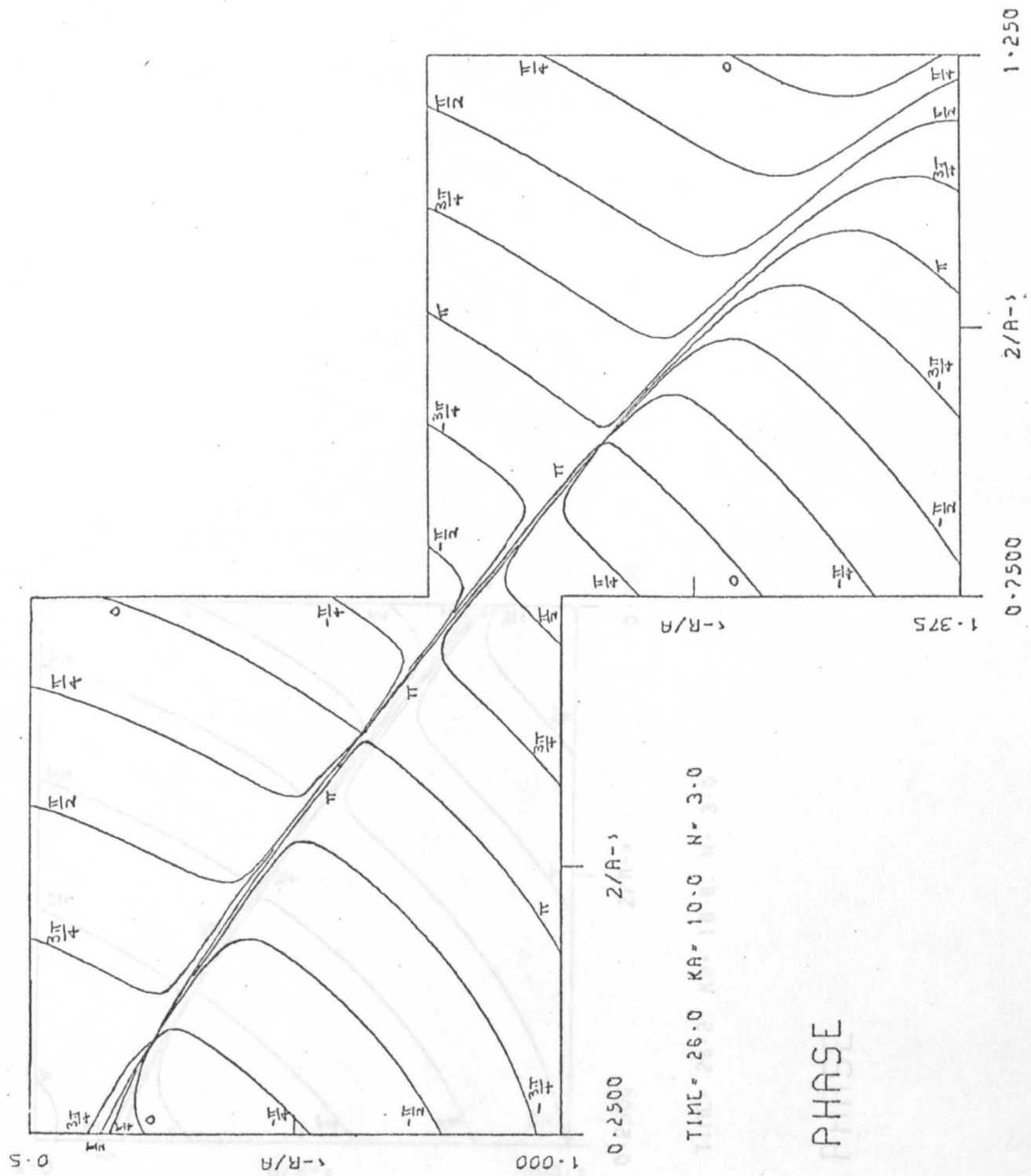


Figure 2.22a Three Dislocations involved in a "Skip":-

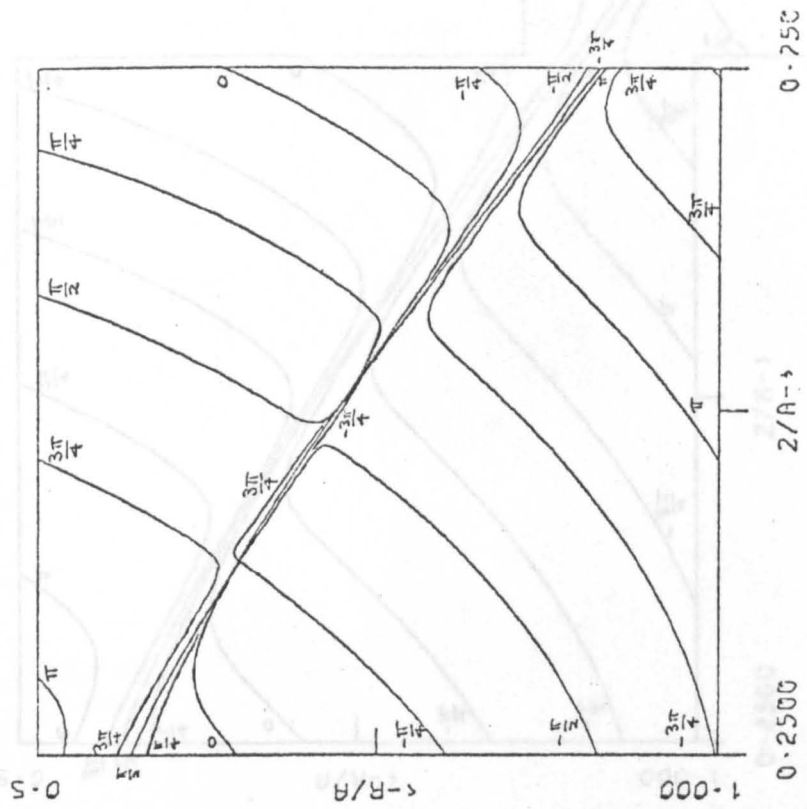


Figure 2.22b A Pair of Opposite Sign Approach and...

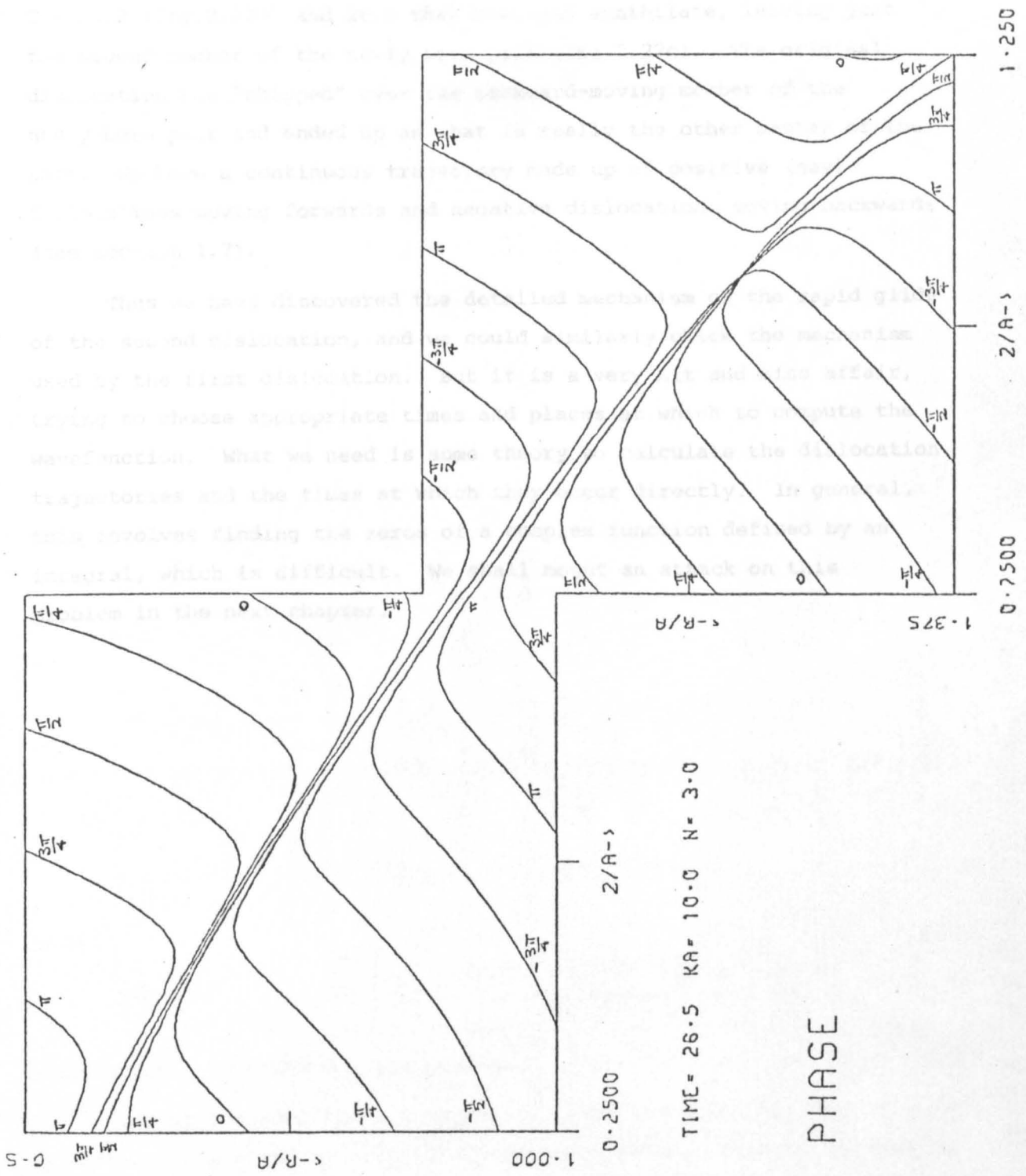


Figure 2.22c ...Annihilate, leaving One Dislocation which appears to have Glided Very Rapidly

This new dislocation has opposite sign to the original, and between T = 26.2 (fig.2.22b) and 26.5 they meet and annihilate, leaving just the second member of the newly born pair (fig.2.22c). The original dislocation has "skipped" over the backward-moving member of the newly-born pair and ended up as what is really the other member of the pair. We have a continuous trajectory made up of positive (say) dislocations moving forwards and negative dislocations moving backwards (see section 1.7).

Thus we have discovered the detailed mechanism of the rapid glide of the second dislocation, and we could similarly check the mechanism used by the first dislocation. But it is a very hit and miss affair, trying to choose appropriate times and places at which to compute the wavefunction. What we need is some theory to calculate the dislocation trajectories and the times at which they occur directly. In general, this involves finding the zeros of a complex function defined by an integral, which is difficult. We shall mount an attack on this problem in the next chapter.

APPENDIX

A2.1 Derivation of Rayleigh's Formula

A general solution of the wave equation may be written in terms of an angular spectrum of plane waves with wave vectors $\underline{k} \equiv (\underline{K},k_z)$ and frequencies ω as

$$\psi(\underline{r},t) = \int_{-\infty}^{\infty} d\omega \int d^3\underline{k}\; e^{i(\underline{k}.\underline{r} - \omega t)}\; A(\underline{k},\omega)$$

subject to $\underline{k}^2 = \underline{K}^2 + k_z^2 = (\omega/c)^2$ and $0 \leq \arg(k_z) < \pi$ (for non-divergent propagation in the half-plane $z > 0$, see fig.2.1 of section 2.1). Therefore

$$A(\underline{k},\omega) = a(\underline{K},\omega)\,\delta(k_z - \sqrt{(\omega/c)^2 - \underline{K}^2})$$

and we may integrate over k_z giving

$$\psi(\underline{r},t) = \int_{-\infty}^{\infty} d\omega \int d^2\underline{K}\; e^{i(\underline{K}.\underline{R} + \sqrt{(\omega/c)^2 - \underline{K}^2}\, z - \omega t)}\; a(\underline{K},\omega) \qquad \text{(A2.1.1)}$$

But on the baffle plane ($z = 0$) we have the boundary condition

$$-\partial\psi/\partial z\big|_{z=0} = G(\underline{R})F'(t)$$

$$\Rightarrow \quad G(\underline{R})F'(t) = -\int_{-\infty}^{\infty} d\omega \int d^2\underline{K} \left[a(\underline{K},\omega)\, i\sqrt{(\omega/c)^2 - \underline{K}^2}\right] e^{i(\underline{K}.\underline{R} - \omega t)}$$

Inverting the Fourier transform and substituting for $a(\underline{K},\omega)$ in (A2.1.1) gives

$$\psi(\underline{r},t) = \int dt_o \int d^2\underline{R}_o\; \chi(\underline{R}-\underline{R}_o,z,t-t_o)\; G(\underline{R}_o)\; F'(t_o) \qquad \text{(A2.1.2)}$$

where

$$\chi(\underline{R},z,t) = \frac{i}{(2\pi)^3}\int_{-\infty}^{\infty} d\omega \int d^2\underline{K}\; \frac{e^{i(\underline{K}.\underline{R} + \sqrt{(\omega/c)^2 - \underline{K}^2}\, z - \omega t)}}{\sqrt{(\omega/c)^2 - \underline{K}^2}} \qquad \text{(A2.1.3)}$$

is the Green function for the problem.

Let us consider the ω integration. The integrand has branch points at $\omega/c = \pm K$ ($K = |\underline{K}|$), around which we must distort the contour of integration. We will consider closing the contour by an infinite semicircle on which the integrand is zero, as shown in fig.A2.1.1. $\chi(\underline{R},z,t)$ itself is the wavefunction produced by the point pulse excitation $G(\underline{R}_o)F'(t_o) = \delta(\underline{R}_o)\,\delta(t_o)$ and causality requires that this be zero everywhere for $t < 0$. To ensure this (for $z = 0$ in particular) we raise the contour along the real axis infinitessimally above the branch points (joined by a branch cut) so that for $t < 0$ the contour

encloses no singularities. Therefore we replace ω in the integrand of A2.1.3 by $\omega_+ = \omega + i\varepsilon$ and eventually take the limit as $\varepsilon \to 0$.

Now that the integrals are well defined, we may put $\underline{K}.\underline{R} = KR\cos\theta$, $d^2\underline{K} = KdKd\theta$ and do the angular integral to give

$$\chi(\underline{R},z,t) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} d\omega\, e^{-i\omega_+ t}\int_0^{\infty} KdK\, ie^{\frac{i\sqrt{(\omega_+/c)^2 - K^2}\, z}{\sqrt{(\omega_+/c)^2 - K^2}}} J_0(KR)$$

where $J_0(KR)$ is a Bessel function (see e.g. Bell (1968)). We can write

$$i\sqrt{(\omega_+/c)^2 - K^2} = \pm\sqrt{K^2 - (\omega_+/c)^2}$$

The contour and branch cut in the plane of $K^2 - (\omega_+/c)^2$ are already imposed on us by our choice in fig.A2.1.1, and are shown in fig.A2.1.2, along with the contour in the plane of $\sqrt{K^2 - (\omega_+/c)^2}$. We see that as the argument $\to +\infty$, the square root $\to -\infty$ i.e. we have been forced to take the negative square root relative to the conventional choice of branch cut along the negative real axis. Then

$$\chi(\underline{R},z,t) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} d\omega\, e^{-i\omega_+ t}\int_0^{\infty} (K^2 + \beta^2)^{-\frac{1}{2}}\, e^{-z(K^2 + \beta^2)^{\frac{1}{2}}} J_0(KR)\, KdK$$

where $\beta = -i\omega_+/c = (\varepsilon - i\omega)/c$ so that Re $\beta > 0$, R and z both real and > 0. We may now apply eqn(24),p9, vol II of Tables of Integral Transforms to give

$$\chi(\underline{R},z,t) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} d\omega\, e^{-i\omega t}\, (R^2 + z^2)^{-\frac{1}{2}}\, e^{i(\omega/c)(R^2 + z^2)^{\frac{1}{2}}}$$

in the limit $\varepsilon \to 0$

$$= \frac{1}{2\pi\rho}\,\delta(t - \rho/c)$$

where $\rho = (\underline{R}^2 + z^2)^{\frac{1}{2}}$. This is the standard result (e.g. see Jackson (1962), p185). Substituting in (A2.1.2) gives our version of Rayleigh's Formula:

$$\psi(\underline{r},t) = \frac{1}{2\pi}\int d^2\underline{R}_0\, \frac{G(\underline{R}_0)\, F'(t - \rho/c)}{\rho}$$

where now $\rho = (|\underline{R} - \underline{R}_0|^2 + z^2)^{\frac{1}{2}}$.

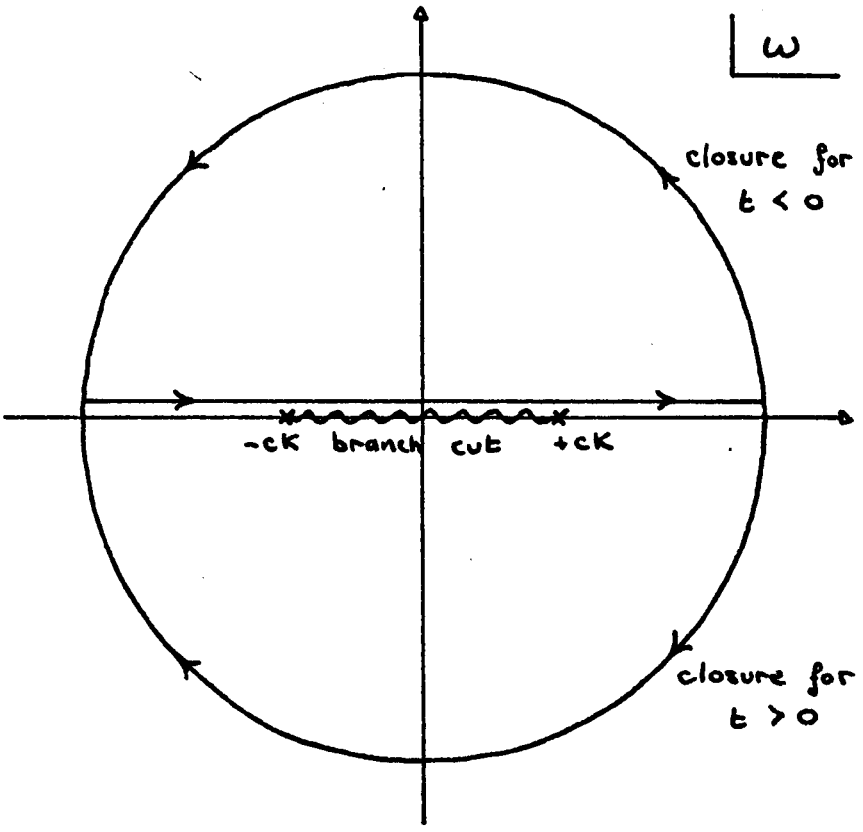


Figure A2.1.1 Contours for the ω Integral

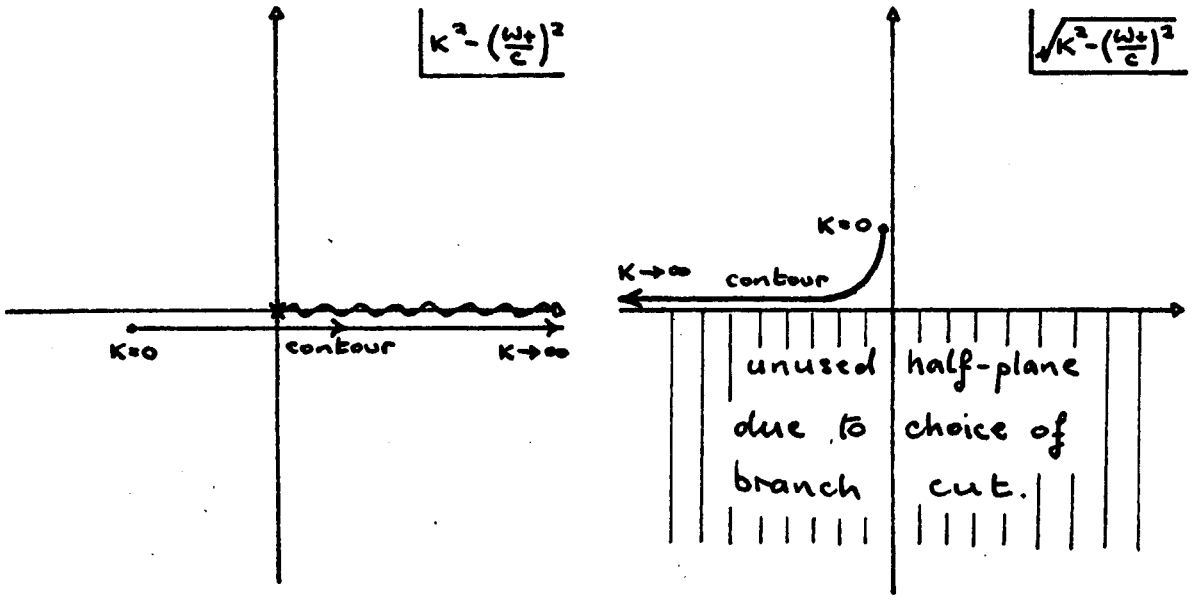


Figure A2.1.2 Contours for the K integral

CHAPTER 3

DISLOCATIONS BY PERTURBATION

3.1 Introduction

In the previous chapter we examined the behaviour of the dislocations occuring in a particular model wavefield, by an "experimental" method. Our apparatus for this experiment was a computer linked to a graph plotter, and using this apparatus we merely examined the spatial wavefield at a sequence of times in the hope that we would find some dislocations. We were guided in our search by intuitive considerations of the continuous wavefield. Our intuition was vindicated by the fact that we did find two dislocations born in the nearfield, one at the only CW null, and running down the "valleys" of the CW amplitude into the far field. But we also found many unexpected phenomena, such as the skip, the pair birth with one of the dislocations moving back onto the piston, and the "incipient" dislocation, about which we still know very little. How many other interesting phenomena did we miss by not choosing the correct times? Looking for dislocations in a wavefield by this method is somewhat akin to looking for a needle moving through a cornfield: it is a hit-or-miss process.

Ideally, we would like to be able to perform one computation to give us the 4-trajectories of all the dislocations in the wavefield. This would tell us the full history of all the dislocations, including the relative signs of all interacting dislocations, as explained in chapter 1. All that it would not tell us is the actual strength of (the dislocations on) each trajectory, and the relative signs of non-interacting dislocations. To find these two facts we would have to compute phases, but we would know exactly where and when to compute them. A half-way step towards this ideal would be to be able to compute only the spatial trajectories (or perhaps only the times).

To solve this problem in general would require finding the zeros of a complex function of 4 variables defined by an arbitrary (diffraction) integral, and I am not aware of any general method for doing this. (We are here allowing dislocations of strength zero, so that a necessary and sufficient condition for a dislocation is that the wave amplitude be zero). However, we are only considering quasimonochromatic pulsed

wavefunctions, so that in some sense our wavefield is only a small deviation from the continuous wavefield at the centre frequency of our pulse. We have already seen how the dislocation trajectory is related to the CW amplitude, so we might hope to be able to derive approximate dislocation 4-trajectories by perturbation of the continuous wavefunction. The question is, can we feasibly make the approximation good enough to be useful?

3.2 General Perturbation Series

By definition, the frequency spectrum of a quasimonochromatic pulse is sharply peaked about its centre frequency, so that the "width" σ of the spectrum is small, and tends to zero in the monochromatic limit of a δ-function. We shall try to develop a perturbation series in σ for the pulsed wavefunction. But by perturbing the width of a δ-function from zero to some finite value, we also "perturb" the height from infinity to some finite value; a "perturbation" which is infinitely large. So we must proceed with care because we are probably building our theory on shifting sand; which may account for some of the peculiar behaviour it exhibits. But then we are hoping to produce nothing out of something; we are not just trying to move the CW nulls a small distance by the perturbation, we are trying to turn a finite amplitude into a zero. The perturbation is only small in terms of the spectral width, not in terms of the amplitude changes it produces!

Suppose we drive the system with the pulse $f(t)e^{i\omega_0 t}$ whose Fourier transform is $\tilde{f}(\omega-\omega_0)$, where

$$\tilde{f}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt\,.$$

If the continuous wavefunction with frequency ω at any point in space and time is $\psi(\underline{r},t,\omega)$, then the pulsed wavefunction is

$$\Psi(\underline{r},t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tilde{f}(\omega-\omega_0)\;\psi(\underline{r},t,\omega)\;d\omega$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tilde{f}(W)\;\psi(\underline{r},t,\omega_0+W)\;dW$$

by putting $W = \omega - \omega_0$. If ψ has phase χ we may write

$$\psi(\underline{r},t,\omega_0+W) = e^{iW\chi_0'}\;P(\underline{r},t,W)$$

where $\chi_0' \equiv \frac{\partial\chi}{\partial\omega}(\omega_0)$ must have the form $t + \bar{\chi}_0'(\underline{r})$. We have factored out of ψ the term with phase varying most rapidly with ω about ω_0, such that the phase of P is stationary at $W = 0$. If $\tilde{f}(W)$ is sufficiently sharply peaked any phase variation of P will not have a large effect. We now expand $P(\underline{r},t,W)$ about $W = 0$, giving

$$\Psi(\underline{r},t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tilde{f}(W)\;e^{iW\chi_0'}\sum_{n=0}^{\infty}\frac{W^n}{n!}P_{(n)}(\underline{r},t,0)\;dW$$

$$= \sum_{n=0}^{\infty} \frac{P_{(n)}(\underline{r},t,0)}{n!} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{f}(W)\, W^n\, e^{iW\chi_o'}\, dW$$

$$= \sum_{n=0}^{\infty} \frac{P_{(n)}(\underline{r},t,0)\; f^{(n)}(\chi_o')}{n!\; i^n} \qquad (3.1)$$

where $P_{(n)}(\underline{r},t,W) \equiv \frac{\partial^n P}{\partial W^n}(\underline{r},t,W)$ and $f^{(n)}(t) \equiv \frac{\partial^n f}{\partial t^n}(t)$. Eqn.(3.1) is an exact series expansion of the pulse wavefunction $\Psi(\underline{r},t)$ at some point $\underline{r}$ and time t, in terms of frequency derivatives evaluated at the pulse centre frequency ω_o of the continuous wavefunction $\psi(\underline{r},t,\omega)$ at that same point and time, and derivatives of the envelope function f(t) modulating the source carrier. The 1/n! factor should ensure that the series converges unless either P or f are particularly badly behaved.

Let us display the pulse length explicitly by replacing f(t) by $f(\sigma t)$. Then for a given function f the actual pulse length is $\propto 1/\sigma$, and the spectrum is

$$\frac{1}{\sigma}\, \tilde{f}\left(\frac{\omega-\omega_o}{\sigma}\right)$$

having width σ (and height $\propto 1/\sigma$). We are interested in pulses with small σ (long pulse length), and our general perturbation series (3.1) becomes

$$\Psi(\underline{r},t) = \sum_{n=0}^{\infty} \frac{P_{(n)}(\underline{r},t,0)\; \sigma^n\; f^{(n)}(\sigma\chi_o')}{n!\; i^n}$$

I am grateful to Dr.M.V.Berry for several suggestions which are incorporated into this formulation.

We expect to be able to make this series converge as rapidly as we like by choosing σ small enough. But note that we must expect this value of σ to vary with the parameters $\underline{r}$, t and ω_o, and we must choose σ small enough to make the series converge sufficiently rapidly everywhere. It is only really feasible to compute frequency derivatives up to second order, therefore we shall truncate the series after the n = 2 term and require σ small enough that the error so incurred is not significant anywhere. If we now factor out of P the constant phase, which must include its time variation, we may write

$$P(\underline{r},t,W) = e^{i\chi_o}\, Q(\underline{r},W)$$

where χ_o must have the form $\omega_o t + \bar{\chi}_o(\underline{r})$, and $Q(\underline{r},0)$ is real. Then

our main formula is

$$\Psi(\underline{r},t) \simeq e^{i\chi_0} \{ Q(\underline{r},0) f(\sigma\chi_0') - i\sigma Q_{(1)}(\underline{r},0) f^{(1)}(\sigma\chi_0') - \frac{\sigma^2}{2} Q_{(2)}(\underline{r},0) f^{(2)}(\sigma\chi_0') \} \tag{3.2}$$

To lowest order

$$\Psi(\underline{r},t) \simeq e^{i\chi_0} Q(\underline{r},0) f(\sigma\chi_0')$$

when the carrier phase is $\chi_0(\underline{r},t) = \omega_0 t + \overline{\chi}_0(\underline{r})$, and at the point $\underline{r}$ in space, the envelope of the pulse as a function of time is $\propto f(\sigma\chi_0'(\underline{r},t))$. So we see that roughly speaking $\chi_0(\underline{r},t)$ describes the propagation of the carrier, and $\chi_0'(\underline{r},t)$ describes the propagation of the envelope (ignoring any dislocations) through the diffraction pattern.

We have written $\psi(\underline{r},t,\omega)$ in the form

$$\psi(\underline{r},t,\omega) = e^{i\chi_0(\underline{r},t)} e^{i(\omega-\omega_0)\chi_0'(\underline{r},t)} Q(\underline{r},\omega-\omega_0) \tag{3.3}$$

where χ_0 and χ_0' have the significance described above. Generally Q is not pure real (except at $\omega = \omega_0$, by definition), but in certain special cases it is. This makes the above decomposition of ψ particularly reasonable, and we might expect the theory to work particularly well in these regions. The condition for a dislocation is that $\Psi(\underline{r},t) = 0$, i.e.

$$Q(\underline{r},0) f(\sigma\chi_0') - i\sigma Q_{(1)}(\underline{r},0) f^{(1)}(\sigma\chi_0') - \frac{\sigma^2}{2} Q_{(2)}(\underline{r},0) f^{(2)}(\sigma\chi_0') = 0 \tag{3.4}$$

and the problem is to solve this generally. We shall always use a real envelope function $f(\sigma t)$, in particular the Gaussian

$$\left.\begin{aligned} f(\sigma t) &= e^{-(\sigma t)^2/2} \\ f^{(1)}(\sigma t) &= -\sigma t\, e^{-(\sigma t)^2/2} \\ f^{(2)}(\sigma t) &= ((\sigma t)^2 - 1)\, e^{-(\sigma t)^2/2} \end{aligned}\right\} \tag{3.5}$$

When Q is also real (3.4) is particularly easy to solve. We shall now apply this theory to the piston radiator discussed in the previous chapter.

3.3 Farfield and Axial Dislocations

If Q is pure real, we can either take it and the phase as analytic functions, so that Q may take negative as well as positive values, or we may take Q as the modulus $|\psi|$, when neither it nor the phase are analytic at Q = 0. We shall take the former interpretation to avoid any non-analyticity, which is probably the main simplification arising from Q being real. Then χ is actually only phase modulo π. Equating real and imaginary parts of (3.4) to zero gives

$$\left.\begin{aligned} Q(\underline{r},0)f(\sigma\chi_o') - \frac{\sigma^2}{2}Q_{(2)}(\underline{r},0)f^{(2)}(\sigma\chi_o') &= 0 \qquad (3.6a)\\ \sigma Q_{(1)}(\underline{r},0)f^{(1)}(\sigma\chi_o') &= 0 \qquad (3.6b)\end{aligned}\right\}$$

For $\sigma \neq 0$, (3.6b) has two solutions

$$f^{(1)}(\sigma\chi_o') = 0 \qquad (3.7a)$$

or $$Q_{(1)}(\underline{r},0) = 0 \qquad (3.7b)$$

Case A: (3.7a) implies the pulse envelope is flat with respect to time, to lowest order i.e. the pulse at that point in space and time is locally "as monochromatic as possible". For the Gaussian envelope (3.5), eqn.(3.6a) is

$$Q(\underline{r},0) - \frac{\sigma^2}{2}((\sigma\chi_o')^2-1)Q_{(2)}(\underline{r},0) = 0 \qquad (3.8)$$

and (3.7a) requires $\chi_o' = 0$. Then (3.8) becomes

$$Q(\underline{r},0) + \frac{\sigma^2}{2}Q_{(2)}(\underline{r},0) = 0 \qquad (3.9)$$

which gives us the spatial position of the dislocation, and $\chi_o' = 0$ tells us the time at which it occurs there (i.e. at the "centre" of the pulse).

Now (3.9) appears to be the equation of a surface, but the condition that Q be real (i.e. Im $Q(\underline{r},0) = 0$) has already restricted us to be on a surface, so that (3.9) actually gives a line (generally). For $\sigma = 0$ this line is the CW null line, so for small σ it must be a line close to the CW null line. The dislocation can only occur at the CW null if $Q_{(2)}(\underline{r},0) = 0$ when $Q(\underline{r},0) = 0$.

Case B: (3.7b) gives the equation of the dislocation line in space, and substituting this into (3.8) gives the time as

$$\chi_0' = \pm \frac{1}{\sigma}\sqrt{\frac{2Q(\underline{r},0)}{\sigma^2 Q_{(2)}(\underline{r},0)} + 1}$$

if this exists. But $Q_{(1)} = 0$ means Q is a maximum or minimum (generally) and we expect Q positive at a maximum and Q negative at a minimum, so that Q and $Q_{(2)}$ have opposite signs. Then for small σ the square root will be imaginary, giving no real solution for χ_0' and no dislocation.

As our first example we consider the far field of the piston radiator. The monochromatic wave function is, from eqn.(2.17),

$$\psi_{Fm}(r,\theta,t) = ie^{i\omega_0(t - r/c)}\, e^{i(\omega-\omega_0)(t - r/c)}\, \frac{ac}{r\sin\theta} J_1\left(\frac{\omega a \sin\theta}{c}\right)$$

Comparison with (3.3) shows that

$$\chi_0 = \pi/2 + \omega_0(t - r/c)$$

$$\chi_0' = t - r/c$$

$$Q(r,\theta,\omega-\omega_0) = J_1\left(\frac{\omega a \sin\theta}{c}\right) \times \text{(term independent of } \omega)$$

Case A gives dislocations at $t - r/c = 0$ and

$$J_1(x) + \frac{x^2}{2(\pi n)^2} J_1''(x) = 0$$

where $x = \frac{\omega_0 a \sin\theta}{c} = K\sin\theta$ and $\sigma = \frac{\omega}{\pi n}$

Using the properties of Bessel functions (e.g. see Bell (1968), p104) this becomes

$$(2(\pi n)^2 + 2 - x^2)J_1(x) - xJ_0(x) = 0 \qquad (3.10)$$

Assuming the dislocations are near to CW nulls, expand x about x_0, where $J_1(x_0) = 0$, by putting $x = x_0 + \delta x$. Then to lowest order

$$\delta x \simeq \frac{x_0}{2(\pi n)^2 + 1 - x_0^2}$$

and $$\delta\theta = \frac{\delta x}{\sqrt{K^2 - x_0^2}}$$

$K = 10$, $n = 3$ gives the following values

x_0	$\delta\theta$	CW null θ_0	dislocation $\theta=\theta_0+\delta\theta$
3·8317	0·15	22·53	22·68
7·0156	0·44	44·55	44·99

(These results were also checked by a direct graphical solution of (3.10), which agreed to the accuracy shown).

Case B gives dislocations at $J_1'(x) = 0$ and

$$t - r/c = \pm \frac{\pi n}{\omega} \sqrt{\frac{2(\pi n)^2 J_1(x) + 1}{x^2 J_1''(x)}}$$

But for $J_1'(x) = 0$, using the properties of Bessel functions, the term inside the square root is

$$\frac{2(\pi n)^2}{1-x^2} + 1 \qquad (3.11)$$

Now $J_1'(x) = 0 \Rightarrow x > 1$ (e.g. see graph of $J_1(x)$ on p134 of Bell). But $x = K \sin\theta \Rightarrow x < K$. Therefore (3.11) is

$$1 - \frac{2(\pi n)^2}{x^2-1} \quad < \quad 1 - \frac{2(\pi n)^2}{K^2-1}$$

If the right hand side is < 0, i.e. if

$$\frac{n}{K} > \frac{1}{\pi\sqrt{2}} = 0.225 \qquad (3.12)$$

then there is no solution for the dislocation. For a long enough pulse this inequality is satisfied, as in our case where $n/K = 0.3$, therefore case B has no solution.

Our perturbation theory predicts two far field dislocations at $t = r/c$ i.e. at the centre of the pulse, and at angles very close to the CW null angles. This agrees with our previous computations, discussed in section 2.7, to the accuracy available from the graphs. To check the accuracy properly, the wavefunction of eqn.(2.23) was computed to an accuracy of 1 in 10^6 for $\tau = 0$, at a sequence of angles closely spaced around the expected zeros. From a graph, the angles of the dislocations were found to be 22.67^0 and 44.88^0 (to 2 DP) (the CW null angles were also checked, and agreed with those quoted). The values of $\delta\theta$ predicted by the perturbation theory are in error by 0.01 in 0.15 (7%) and 0.1 in 0.4 (25%) respectively. But our pulse is only just above the threshold (3.12) for spurious solutions, and

presumably for longer pulses the accuracy would improve. We have, however, predicted the time exactly!

Our second example is the axial field of the piston radiator. The monochromatic wavefunction is, from equation (2.19)

$$\Psi_m(0,Z,\omega t) = 2ie^{i\omega_0(t - \tilde{Z}a/c)}\, e^{i(\omega-\omega_0)(t - \tilde{Z}a/c)} \sin \frac{\omega a}{c}\left(\frac{Z_1-Z}{2}\right)$$

where $Z = z/a$, $Z_1 = \sqrt{Z^2+1}$ and $\tilde{Z} = (Z_1+Z)/2$.
Comparison with (3.3) shows that

$$\chi_o = \pi/2 + \omega_o(t - \tilde{Z}a/c)$$

$$\chi_o' = t - \tilde{Z}a/c$$

$$Q(Z,\omega-\omega_o) = \sin \frac{\omega a}{c}\left(\frac{Z_1-Z}{2}\right) \times \text{constant}$$

Case A gives dislocations at $t - \tilde{Z}a/c = 0$, or $T - K\tilde{Z} = 0$
where $T = \omega_o t$, $K = \omega_o a/c$ and

$$\sin K\left(\frac{Z_1-Z}{2}\right)\left\{1 - \frac{1}{2(\pi n)^2} K^2\left(\frac{Z_1-Z}{2}\right)^2\right\} = 0$$

Now $(Z_1-Z)/2$ decreases monotonically from ½ to 0 as Z increases from 0 to ∞, so if

$$\frac{K^2}{2(\pi n)^2} < 4 \quad \text{i.e.} \quad \frac{n}{K} > \frac{1}{2\pi\sqrt{2}} = 0.113 \qquad (3.13)$$

the only solution is $\sin K(Z_1-Z)/2 = 0$ which is precisely the CW null. Condition (3.13) is satisfied by a long enough pulse, which ours is.

Case B gives dislocations at

$$\frac{a}{c}\left(\frac{Z_1-Z}{2}\right) \cos K\left(\frac{Z_1-Z}{2}\right) = 0$$

We must take $\cos K(Z_1-Z)/2 = 0$ because $(Z_1-Z)/2 \neq 0$ (except at infinity). Then $\sin K(Z_1-Z)/2 \neq 0$, so the time is given by

$$t - \tilde{Z}a/c = \pm \frac{\pi n}{\omega} \sqrt{\frac{-2(\pi n)^2}{(K(Z_1-Z)/2)^2} + 1}$$

Now $((Z_1-Z)/2)^2 < 1/4$

$$\Rightarrow \quad 1 - \frac{2(\pi n)^2}{(K(Z_1-Z)/2)^2} < 1 - \frac{8(\pi n)^2}{K^2}$$

Then if condition (3.13) is satisfied, the above right hand side < 0 and there are no solutions for the dislocations.

Our perturbation theory predicts one dislocation exactly at the CW null, at time $T = K\tilde{Z} = 7.958$ for $K = 10$. This again agrees with our previous computations, discussed in section 2.7, to the accuracy available from the graphs. But we can check this result analytically. Suppose we drive the piston with the pulse

$$F(t) = f(\omega t)\, e^{i\omega t},$$

then the axial wavefunction is

$$\Psi(0,Z,T) = f(T - KZ)\, e^{i(T - KZ)} - f(T - KZ_1)\, e^{i(T - KZ_1)} \qquad (3.14)$$

We require $\Psi = 0$. If $f(T)$ is real and > 0 for all T, then the amplitudes and phases of the two terms in (3.14) must be equal. The phase condition gives

$$T - KZ = T - KZ_1 + 2n\pi$$

$$\Rightarrow \quad K(Z_1 - Z)/2 = n\pi$$

which is the condition for a CW null (see eqn. (2.20)) (because the amplitudes are equal if $f(T) = 1$ for all T). The amplitude condition is

$$f(T - KZ) = f(T - KZ_1) \qquad (3.15)$$

Suppose $f(T)$ is symmetric and monotonically decreasing away from $T = 0$, then (3.15) implies

$$T - KZ = T - KZ_1$$

or $$T - KZ = -(T - KZ_1).$$

Since $Z_1 \neq Z$ (except at infinity) we must take the second equation, giving

$$T = K(Z + Z_1)/2 = K\tilde{Z}$$

We see that for <u>any</u> pulse whose envelope function is real, positive definite, symmetric and monotonic decreasing away from its centre, axial dislocations occur exactly at the CW axial nulls at time $T = K\tilde{Z}$, which is the mean travel time of the wave from the centre and edge of the piston, and nowhere else. This is independent of the detailed pulse shape or pulse length, and our Gaussian model is just one of many possible envelopes satisfying the above conditions.

For a long enough Gaussian pulse, the perturbation theory gives the axial dislocations exactly. Note that

$$\frac{\pi n}{K} = \frac{\lambda n}{2a} = \frac{\text{length of pulse in space}}{\text{diameter of piston}},$$

and if this is much less than 1, then the main pulse (if there is one) and the replica pulses from opposite edges of the piston will be well separated, rather than merging into a single pulse as the theory tacitly assumes, which may well create dislocations between them. The condition that $\pi n/K > 1/\sqrt{2}$ avoids this complication in the far field and on the axis.

3.4 General Solution of the Perturbation Equation

To solve equation (3.4) we must resolve it into its real and imaginary parts. Just taking the real and imaginary parts of Q is not very useful, so we write

$$Q(W) = R(W)\, e^{i\{\chi_o'' W^2/2 + O(W^3)\}} \tag{3.16}$$

where the $\underline{r}$ dependence is not shown explicitly, and R and χ_o'' are real. Then (3.4) becomes

$$\{R(0)f(\sigma\chi_o') - i\sigma R_{(1)}(0)f^{(1)}(\sigma\chi_o') - \frac{\sigma^2}{2}\{R_{(2)}(0)+iR(0)\chi_o''\}f^{(2)}(\sigma\chi_o')\}\, e^{i\{\cdots\}} = 0$$

and separating real and imaginary parts gives

$$\left.\begin{aligned} R(0)f(\sigma\chi_o') - \frac{\sigma^2}{2}R_{(2)}(0)f^{(2)}(\sigma\chi_o') &= 0 \qquad (3.17a)\\ -\sigma\{R_{(1)}(0)f^{(1)}(\sigma\chi_o') + \frac{\sigma}{2}R(0)\chi_o''f^{(2)}(\sigma\chi_o')\} &= 0 \qquad (3.17b)\end{aligned}\right\}$$

When $\chi_o'' = 0$ these equations reduce to (3.6), of course.

In general, the only way we can make the decomposition (3.16) is by insisting that

$$R(W) \geqslant 0 \quad \text{i.e.} \quad R = |\psi|$$

Then the phase, and hence our equations (3.17), are well defined everywhere except where R = 0, i.e. at CW nulls. We consider this exceptional case first. Equations (3.17) give the 4-trajectory of the dislocations arising in the wavefield: 2 equations in 4 variables define a 2-surface. If we eliminate t between them we have the equation of the real trajectory surface in 3-space. Let us consider a curve lying in the trajectory surface, such that the CW amplitude is zero at isolated points on the curve. We consider the restriction of all our functions to this curve. Then we may take R to be a real and analytic function of distance along the curve, which may be negative, so that the phase is also analytic (see appendix A3.1). This is analogous to what we have already done along the axis of the piston radiator. Now equations (3.17) are well-defined everywhere along our curve, and we can find the conditions necessary for such a curve to pass through a CW null. R(0) = 0 in eqns.(3.17) implies

$$\left.\begin{aligned} -\frac{\sigma^2}{2}R_{(2)}(0)f^{(2)}(\sigma\chi_o') &= 0 \qquad (3.18a)\\ -\sigma\, R_{(1)}(0)f^{(1)}(\sigma\chi_o') &= 0 \qquad (3.18b)\end{aligned}\right\}$$

Generically $R_{(1)}(0) \neq 0$ at $R(0) = 0$, therefore (3.18b) requires $f^{(1)}(\sigma\chi_0') = 0$. Then generically $f^{(2)}(\sigma\chi_0') \neq 0$, so (3.18a) requires $R_{(2)}(0) = 0$. Therefore for a pulse with no stationary inflexions, (such as our Gaussian) the dislocation trajectory only passes through CW nulls, $R(0) = 0$, where also $R_{(2)}(0) = 0$. This is likely to occur only in very special situations, such as the very high symmetry position of the axis of the circular piston radiator, where $R \sim \sin$. It does not occur in the far field where $R \sim J_1$. We suggest that the dislocation trajectory does not pass through a CW null in a general position, although for small σ it may pass very close to it.

We have ascertained the condition for a dislocation to pass through a CW null, and that if it does so the time will be given by $f^{(1)}(\sigma\chi_0') = 0$, i.e. the dislocation will pass through the CW null at its "most monochromatic" time. We can now exclude the CW nulls, and regard R as the amplitude or modulus M of the continuous wavefunction, and only consider points in space where $M > 0$. To make this change of view explicit we rewrite (3.17) as

$$\left.\begin{aligned} Mf(\sigma\chi_0') - \frac{\sigma^2}{2}M_2 f^{(2)}(\sigma\chi_0') &= 0 \qquad (3.19a)\\ \sigma\{M_1 f^{(1)}(\sigma\chi_0') + \frac{\sigma M}{2}\chi_0'' f^{(2)}(\sigma\chi_0')\} &= 0 \qquad (3.19b)\end{aligned}\right\}$$

where $M_r \equiv R_{(r)}(0)$. We shall make some general comments about these equations first. M is independent of σ, and everywhere greater than 0, and we expect $M_2\sigma^2$ to be smaller in magnitude than M. We only consider positive definite pulses, therefore for very small σ a solution of (3.19a) requires $M_2 f^{(2)}(\sigma\chi_0') > 0$ and $f^{(2)}/f$ fairly large. For our Gaussian model pulse $f^{(2)}/f = (\sigma\chi_0') - 1$ from (3.5), and this can only have a magnitude greater than 1 if it is positive. Therefore, in a typical region of space, we expect dislocations only for $M_2 > 0$ and $f^{(2)} > 0$, i.e. near frequency minima of the CW amplitude, and for times outside the inflexions of the pulse envelope.

Before we can proceed to solve equations (3.19) we must derive the modulus and phase terms from the continuous wavefunction, which is what we actually know (usually as a diffraction integral). We write

$$\psi(\underline{r},0,\omega) = R(\underline{r},\omega) + iI(\underline{r},\omega) = M(\underline{r},\omega)\, e^{i\bar{\chi}(\underline{r},\omega)}$$

Then using the notation $X_\omega \equiv \partial X/\partial\omega$ we have

$$
\left.
\begin{aligned}
M^2 &= R^2 + I^2 \\
M\,M_\omega &= R\,R_\omega + I\,I_\omega \\
M\,M_{\omega\omega} &= R\,R_{\omega\omega} + I\,I_{\omega\omega} + (R_\omega)^2 + (I_\omega)^2 - (M_\omega)^2 \\
M^2\bar{\chi}_\omega &= R\,I_\omega - I\,R_\omega \\
M^2\bar{\chi}_{\omega\omega} &= R\,I_{\omega\omega} - I\,R_{\omega\omega} - 2M\,M_\omega\,\bar{\chi}_\omega
\end{aligned}
\right\} \qquad (3.20)
$$

If we can compute ψ, ψ_ω and $\psi_{\omega\omega}$ then we can compute all the terms required in (3.19). In principle each of these terms is an analytical function of its arguments, but in practice this analytical form is so complicated as to be useless. The terms in (3.19) are, in terms of those above:

$$
\begin{aligned}
M &= M(\underline{r},\omega_o) \\
M_1 &= M_\omega(\underline{r},\omega_o) \\
M_2 &= M_{\omega\omega}(\underline{r},\omega_o) \\
\chi_o' &= t + \bar{\chi}_\omega(\underline{r},\omega_o) \\
\chi_o'' &= \bar{\chi}_{\omega\omega}(\underline{r},\omega_o)
\end{aligned}
$$

and we regard these as being numerically defined function of space $\underline{r}$. We could compute them at any required point, but what we actually do is compute arrays of all the required functions once for all at lattice points throughout the region of space in which we are interested. We then interpolate between these lattice points.

The pulse function f(t) is an analytical function of its argument, so equations (3.19) have the status of a pair of functions, which are numerical in space and analytical in time, equated to zero. The way to solve these equations depends on the form in which we would like the results. In a truly three-dimensional problem, displaying the results would be difficult. But we are only currently interested in problems which are "effectively two-dimensional", when the shape of the dislocation line is known (to be a straight line for translation symmetry, a circle for rotation symmetry), and the whole dislocation line is represented by a point in our two-dimensional space. Then the 4-trajectory reduces to a line in 3-space (instead of a 2-surface in 4-space), which can be visualized.

We shall plot the projection into real space, which is the real

trajectory curve in two-dimensions, and the time as a function of distance along this curve from some arbitrary origin, for each branch of the trajectory. (We could then stick each time plot along its trajectory, normal to the real plane, and thereby construct the full trajectory curve in our three-dimensional space-time.) Since time appears analytically in (3.19), we should be able to solve for it analytically in terms of numerical functions of position along the trajectory, which can be easily plotted. We should also be able to eliminate time giving the trajectory as the solution of

"numerical function" = 0,

which can be easily plotted using a contouring routine.

3.5 A First Approximation

Before we attempt to solve equations (3.19) in their full form, let us consider a much simpler first approximation solution. To get anything other than the CW nulls we must keep both terms of (3.19a). But the lowest order solution of (3.19b) may be a good approximation for small σ, viz.

$$M_1 f^{(1)}(\sigma\chi_o') = 0 \tag{3.21}$$

Let us return to our model of the piston radiator, and rewrite (3.19) in terms of our dimensionless variables K and n where

$$K = ka \quad \text{and} \quad \sigma = \frac{\omega_o}{\pi n} \qquad (T = \omega_o t)$$

to give

$$Mf - \frac{1}{2(\pi n)^2} M_{KK} f^{(2)} = 0 \tag{3.22a}$$

$$M_K f^{(1)} + \left\{ \frac{1}{2(\pi n)} M\chi_{KK} f^{(2)} \right\} = 0 \tag{3.22b}$$

where the argument of f is $(T + \chi_K)/\pi n$, and

$$M_K = K\, \partial M/\partial K$$

$$M_{KK} = K^2\, \partial^2 M/\partial K^2$$

$$\chi_K = K\, \partial\bar{\chi}/\partial K$$

$$\chi_{KK} = K^2\, \partial^2\bar{\chi}/\partial K^2$$

We display in figs.3.1a to e contours of M, M_K, $M\,M_{KK}$, χ_K, $M^2\,\chi_{KK}$ respectively (the second derivatives being in the form of eqn.(3.20), as originally derived from the wavefunction), plotted for K = 10. (This means that as displayed the first derivatives are multiplied by 10 and the second derivatives by 100.)

Our first approximation equation (3.21) becomes (3.22b) without the term in braces, and its solution is either $M_K = 0$ or $f^{(1)} = 0$. For our Gaussian pulse, the latter implies $T + \chi_K = 0$ gives the times, and then equation (3.22a) gives the trajectory as

$$M + \frac{1}{2(\pi n)^2} M_{KK} = 0$$

(since $f^{(2)}(0) = -f(0) \neq 0$ from (3.5)). This requires $M_{KK} < 0$, but we have already deduced that we generally expect $M_{KK} > 0$ along the

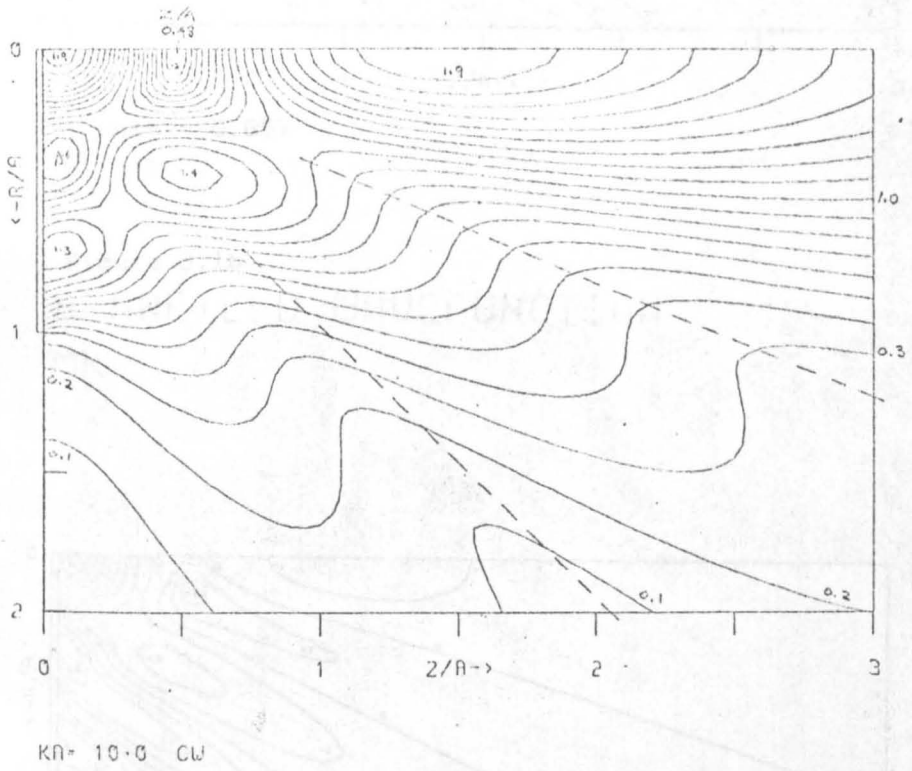
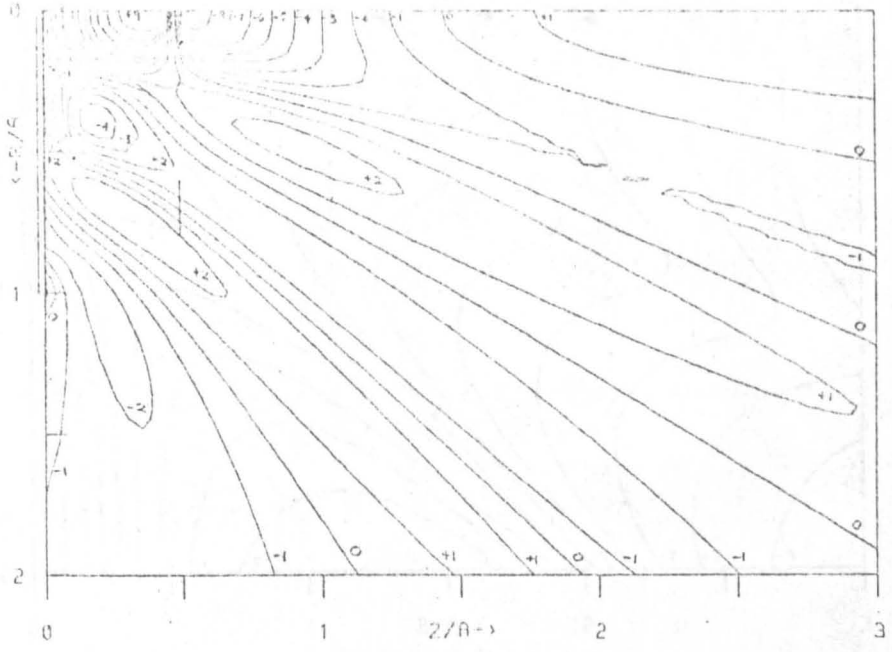


Figure 3.1a

NEARFIELD WAVEFUNCTION
CONTOURS OF AMPLITUDE in steps of 0.1

Dashed lines represent farfield null angles.



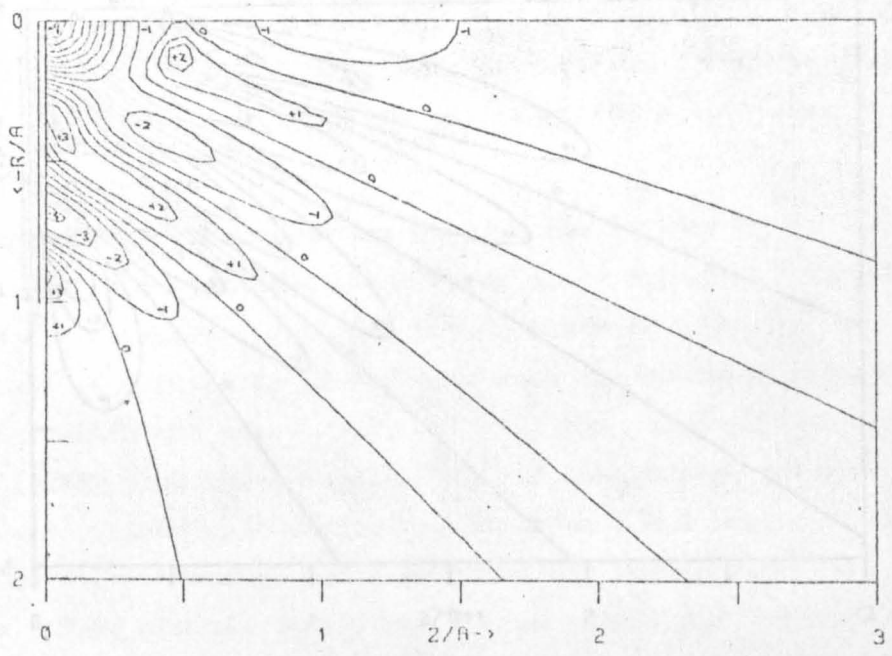
1 unit = 0.657

KA = 10.0

Figure 3.1b

NEARFIELD WAVEFUNCTION

MK



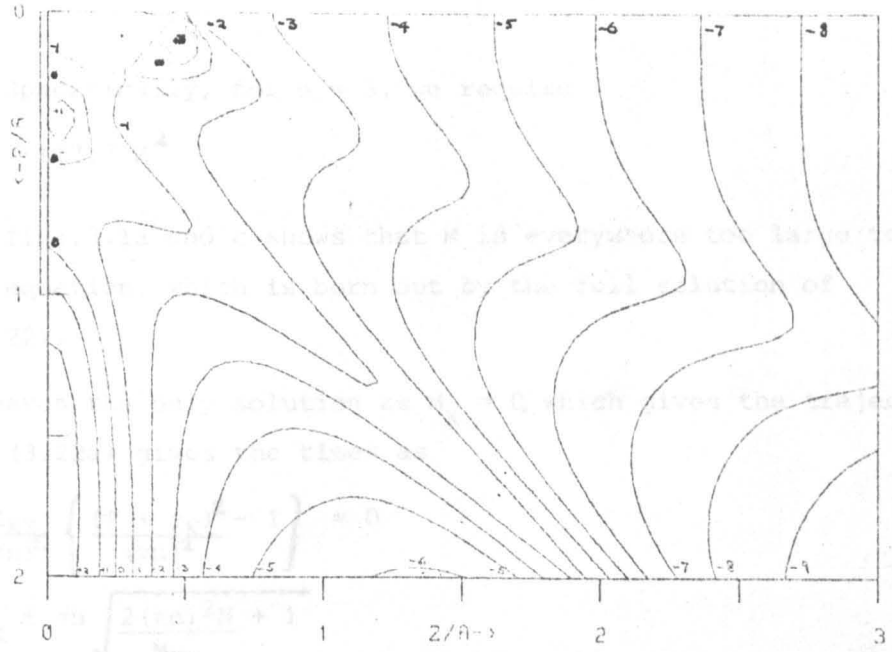
1 unit = 9.195

KA = 10.0

Figure 3.1c

NEARFIELD WAVEFUNCTION

MMKK



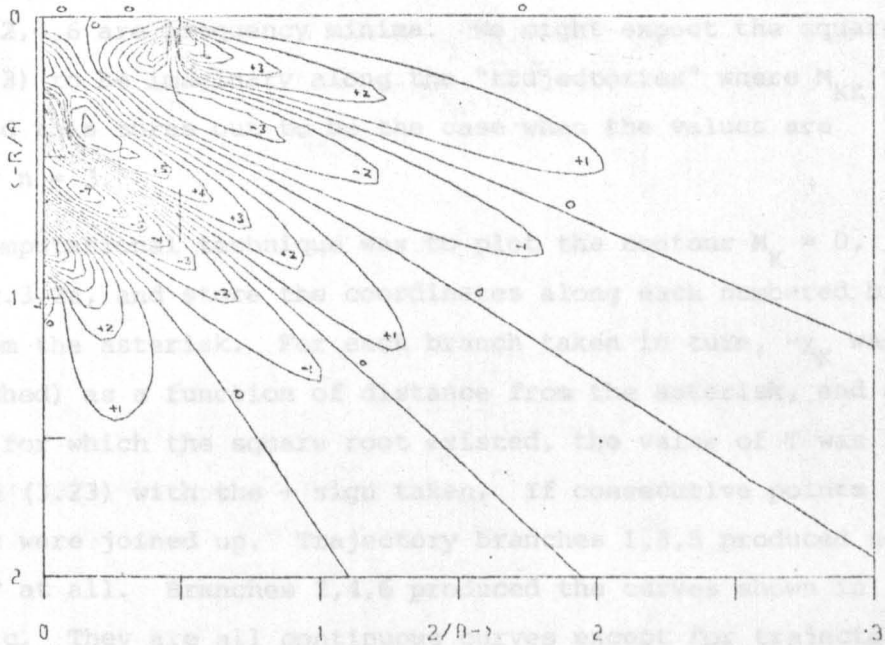
1 unit = 3.505

KA = 10.0

Figure 3.1d

NEARFIELD WAVEFUNCTION

χ_K



1 unit = 4.052

KA = 10.0

Figure 3.1e

NEARFIELD WAVEFUNCTION

$M^2\chi_{KK}$

trajectory. Specifically, for n = 3, we require

$$M\, M_{KK} = -177.7\, M^2$$

Reference to figs.3.1a and c shows that M is everywhere too large to satisfy this equation, which is born out by the full solution of equations (3.22).

This leaves the only solution as $M_K = 0$, which gives the trajectory, and then eqn.(3.22a) gives the times as

$$M - \frac{M_{KK}}{2(\pi n)^2}\left\{\frac{(T + \chi_K)^2}{(\pi n)^2} - 1\right\} = 0$$

$$\Rightarrow \quad T = -\chi_K \pm \pi n \sqrt{\frac{2(\pi n)^2 M}{M_{KK}} + 1} \qquad (3.23)$$

This appears to give dislocations symmetrically in the head and tail of the pulse, and this will occur for any symmetrical pulse in this lowest order approximation. We shall choose the + sign because we know that the dislocations actually occur in the tail, and hope that the full solution of (3.22) will lend some theoretical justification for this. The lines satisfying $M_K = 0$ are visible on the contour plot (fig.3.1b) but we plot them on their own for clarity in fig.3.2. There are 6 lines. Reference to fig.3.1c shows that lines 1,3,5 are frequency maxima, while lines 2,4,6 are frequency minima. We might expect the square root in (3.23) to be imaginary along the "trajectories" where M_{KK} is negative, and this turns out to be the case when the values are computed for n = 3.

The computational technique was to plot the contour $M_K = 0$, shown in fig.3.2a, and store the coordinates along each numbered branch starting from the asterisk. For each branch taken in turn, $-\chi_K$ was plotted (dashed) as a function of distance from the asterisk, and at every point for which the square root existed, the value of T was plotted from (3.23) with the + sign taken. If consecutive points existed they were joined up. Trajectory branches 1,3,5 produced no points for T at all. Branches 2,4,6 produced the curves shown in figs.3.3a,b,c. They are all continuous curves except for trajectory 2 near the CW null. The jaggedness is almost certainly due to the crude linear interpolation used, and could no doubt be improved if the theory was thought to merit it.

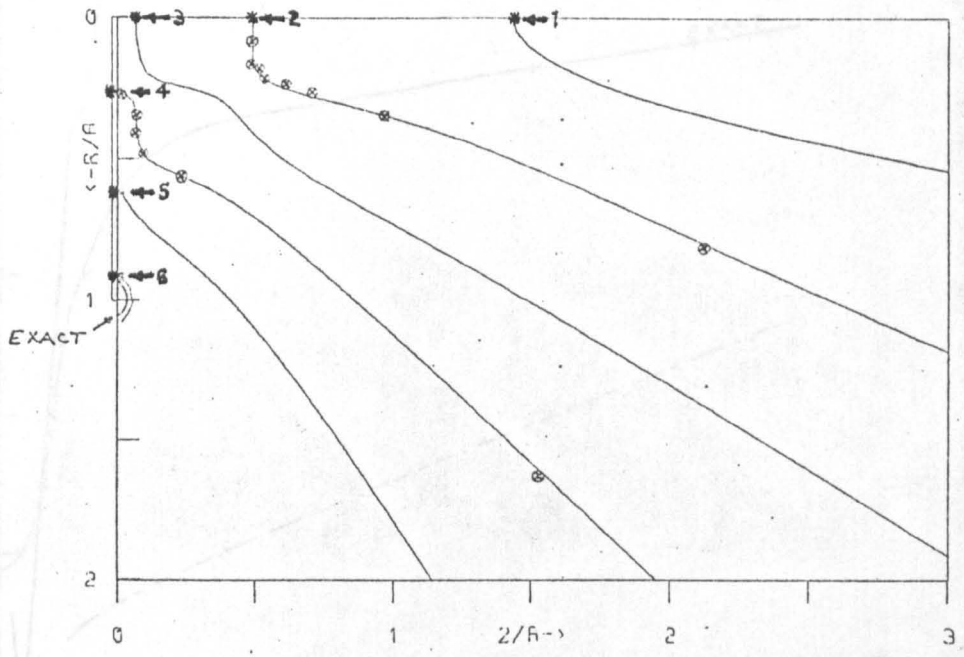
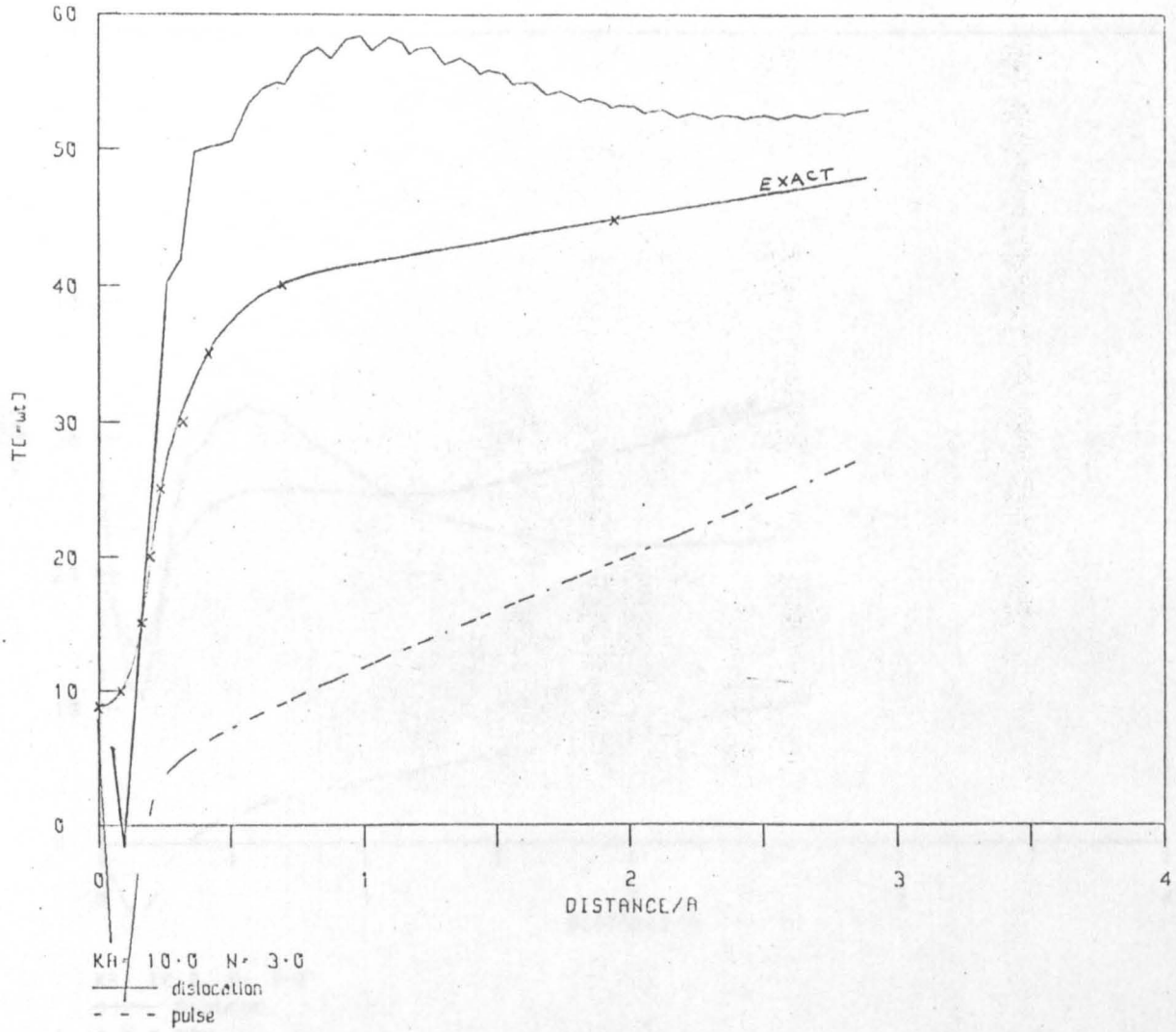


Figure 3.2

NEARFIELD WAVEFUNCTION
DISLOCATION TRAJECTORY

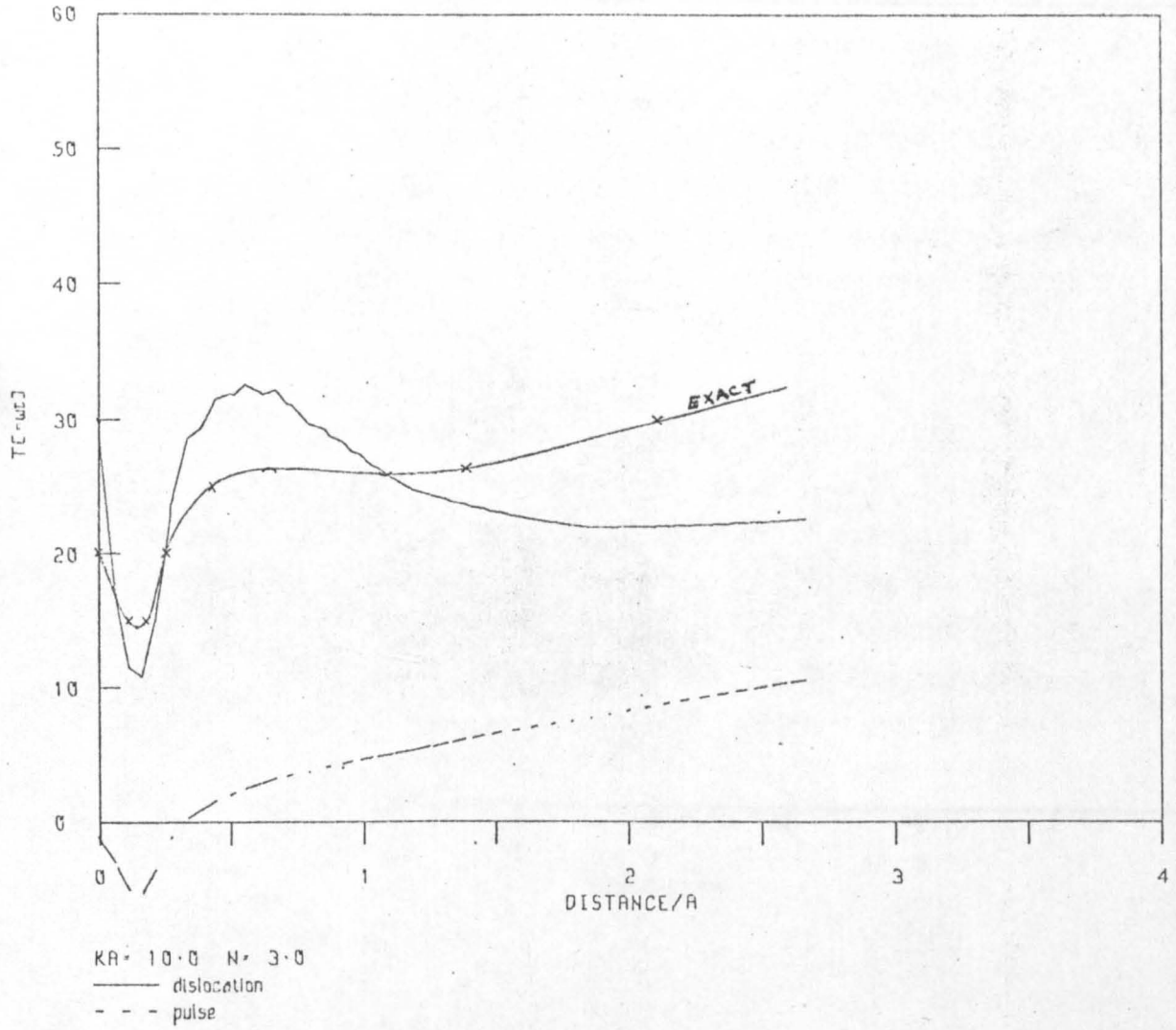
(a) Zero Contour of M_K ———

(b) Computed Points on Exact Trajectory ⊗
and Exact Incipient Trajectory



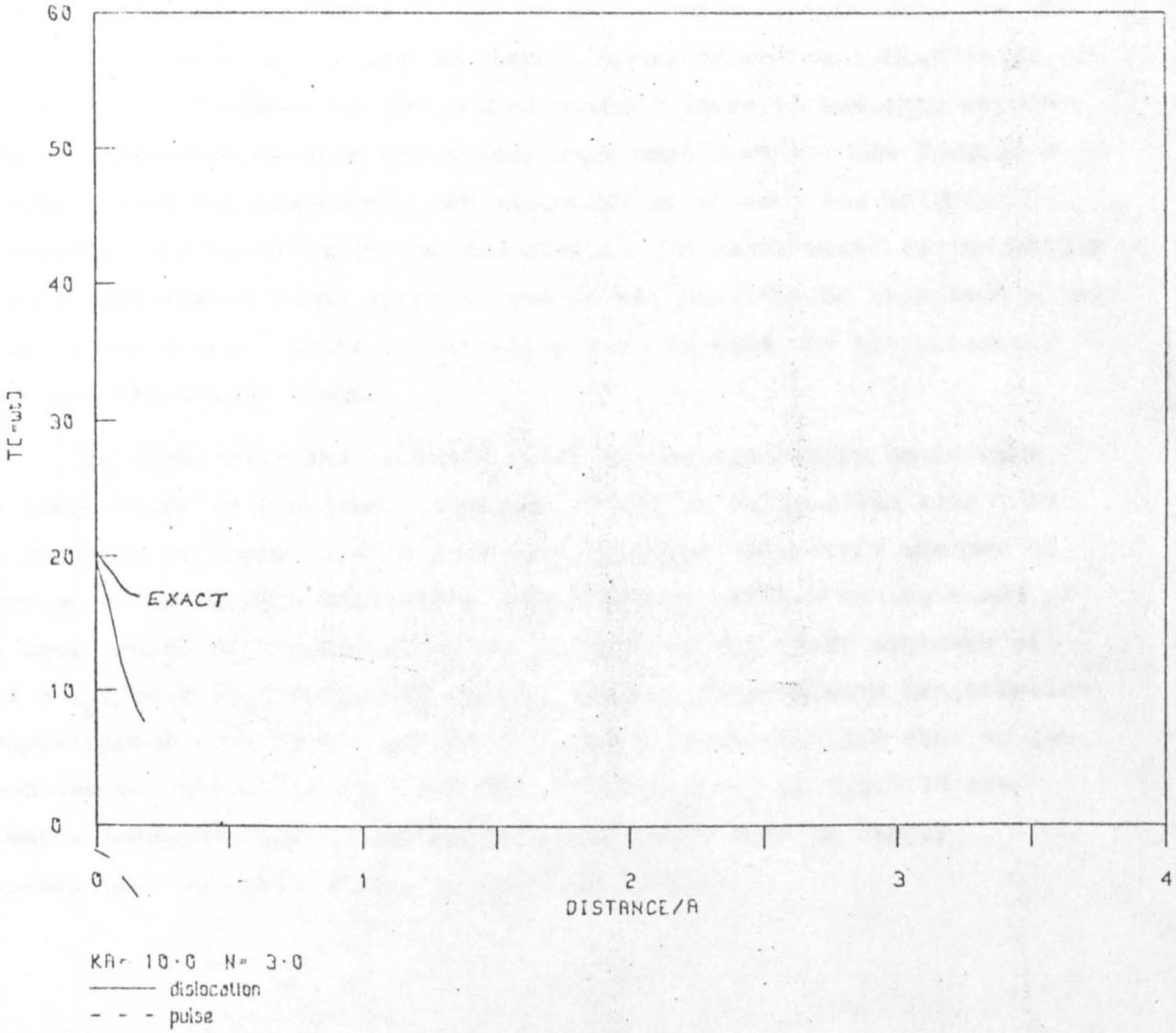
NEARFIELD WAVEFUNCTION
DISLOCATION TIMES - TRAJECTORY 2

Figure 3.3a



NEARFIELD WAVEFUNCTION
DISLOCATION TIMES - TRAJECTORY 4

Figure 3.3b



NEARFIELD WAVEFUNCTION
DISLOCATION TIMES - TRAJECTORY 6

Figure 3.3c

Only the frequency minima of M, curves 2,4 & 6 in fig.3.2a, are dislocation trajectories. These compare very well with the "experimental" trajectories found in the previous chapter, and reproduced for comparison in fig.3.2b. Agreement is within "experimental error", this being the error incurred in plotting the individual dislocations and then tracing the trajectories to produce fig.3.2b. The "experimental" times are also plotted on figures 3.3, the crosses showing the points actually known. Here the agreement is not so good, and such subtleties as the skip along trajectory 4 are completely wrong (just how subtle it is, is shown by the flatness of the "experimental" curve in the skip region). This is probably because the actual wave amplitude has the form of a valley along the trajectory, whose bottom is at very low height, becoming zero actually at the dislocation. A small error in the height could then make a large error in the actual position of this zero along the valley floor. Hence it is fairly easy to find the trajectories, but not the actual times.

We might hope that a small error in the trajectory could make a large error in the time. Equation (3.23) actually gives time T as a function of space, and it must give the time accurately whether or not we solve (3.22b) accurately. To judge the effect on the times of a small shift in the trajectories, we plot in fig.(3.4) contours of $(T + \chi_K)/\pi n = \sigma\chi_o'$ (where it exists) with our approximate trajectories superimposed upon them. We see that the trajectories lie near to the bottoms of valleys of $\sigma\chi_o'$, and the contours of χ_K in fig.3.1d are nearly perpendicular to the trajectories, so T will be fairly insensitive to small shifts in the trajectories.

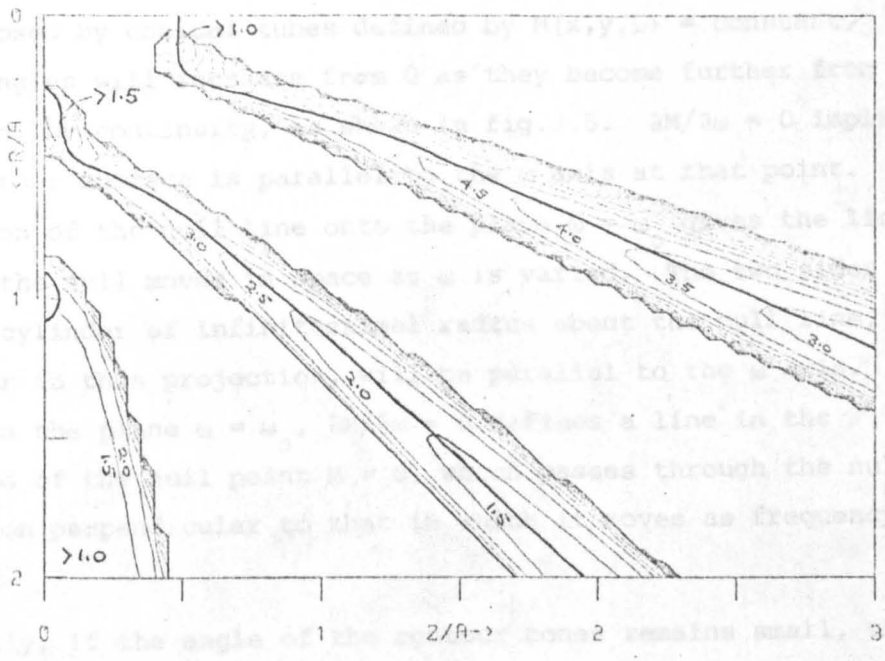


Figure 3.4

NEARFIELD WAVEFUNCTION

Contours of $(T + \chi_K)/\pi n = \sigma\chi_o'$ (from equation (3.23)) with trajectories superimposed

3.6 The Dislocation Trajectories

We have seen that the dislocation trajectories are given to a good approximation by $\partial M/\partial\omega = 0$, which appears to pass through the axial CW null where $M = 0$. We shall show that every CW null line $M(\underline{r},\omega_o) = 0$ (which actually represents 2 equations in general) has a surface $\frac{\partial M}{\partial\omega}(\underline{r},\omega_o) = 0$ passing through it. Consider a plane in space, the xy plane say. Then the contour surfaces of $M(x,y,\omega)$ are two-dimensional surfaces in this three-dimensional space. Generically $M(x,y,\omega) = 0$ defines a line at some angle to the ω axis. This line will be enclosed by conical tubes defined by $M(x,y,\omega) =$ constant, whose semi-angles will increase from 0 as they become further from the null line, by continuity, as shown in fig.3.5. $\partial M/\partial\omega = 0$ implies that the contour surface is parallel to the ω axis at that point. The projection of the null line onto the plane $\omega = \omega_o$ gives the line along which the null moves in space as ω is varied. The two sides of the contour cylinder of infinitessimal radius about the null line, perpendicular to this projection, will be parallel to the ω axis. Therefore, in the plane $\omega = \omega_o$, $\partial M/\partial\omega = 0$ defines a line in the neighbourhood of the null point $M = 0$, which passes through the null in a direction perpendicular to that in which it moves as frequency is varied.

Globally, if the angle of the contour cones remains small, the line will continue away from the null on either side (fig.3.5a). But if the angle of the cones increases, there will be an M contour whose semi-angle equals the inclination of the null line to the ω axis, and outside which the contours cannot be parallel to the ω axis. Therefore the two ends of the line $\partial M/\partial\omega = 0$ must join up, where a generator of the M-contour cone is parallel to the ω axis, into a closed loop (fig. 3.5b). If the null did not move with ω, this loop would degenerate into a point at the null, giving a dislocation rigidly fixed in space, but this is highly unlikely.

By considering a family of parallel planes like that taken as our xy plane, we see that $\partial M/\partial\omega = 0$ represents a natural extension of the CW null lines $M = 0$ into surfaces. These surfaces will be very close to the trajectory surfaces of the dislocations produced by an arbitrary quasimonochromatic pulse, whose centre frequency is the CW frequency.

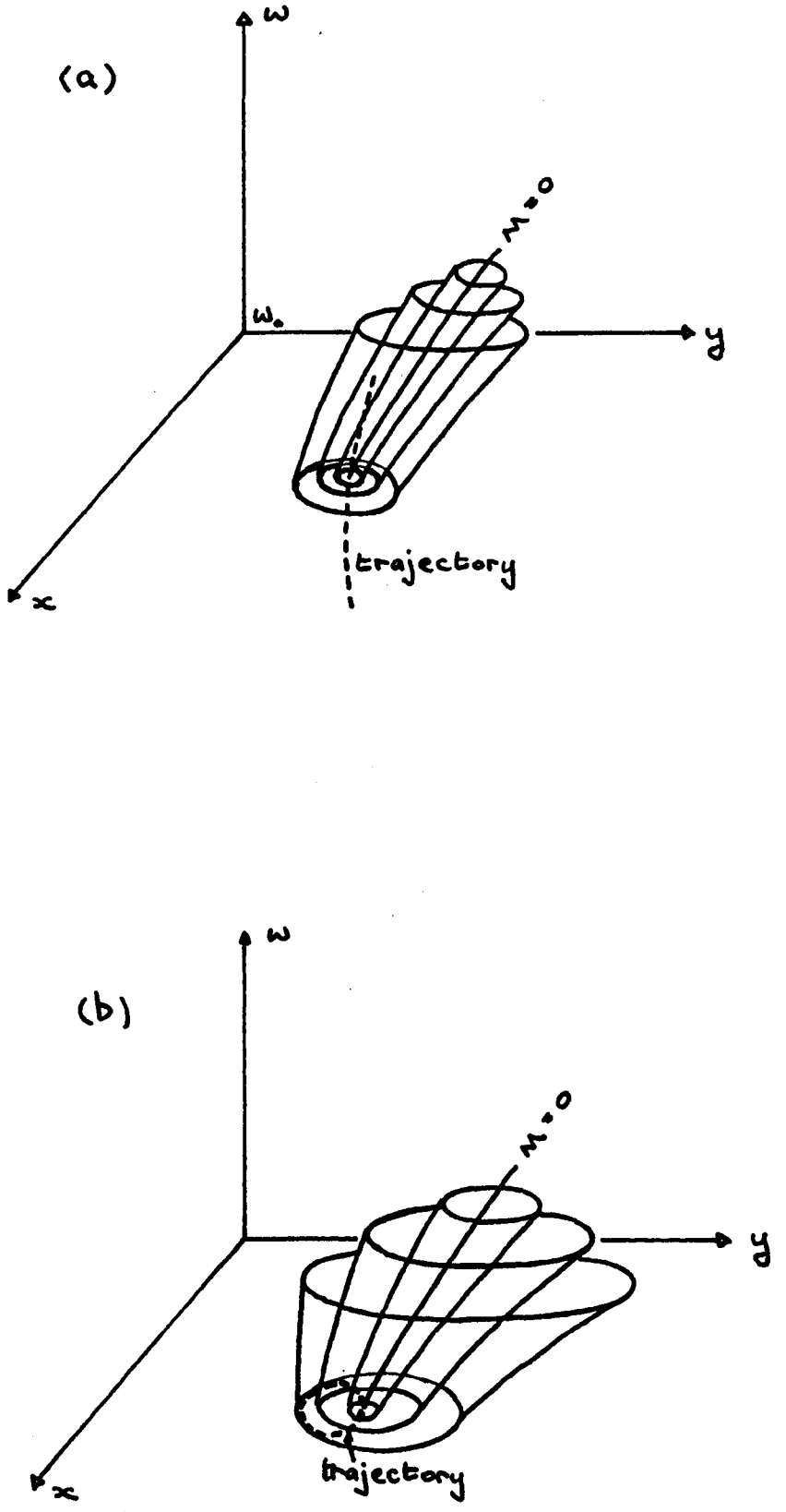


Figure 3.5 Contours of M(x,y,ω)

In the limit of an infinitely long pulse, we expect the trajectory surface to somehow shrink back to the CW null lines. The way that this happens can be seen from eqn.(3.23). As $n \to \infty$, $T \to \infty$ and the dislocations recede into the tail of the pulse, never to appear at any finite time, except where $M = 0$ when the equations break down. Then eqns.(3.17) tell us that in the limit $\sigma = 0$ $(n = \infty)$, $R = 0$ $(M = 0)$ is a solution for all time. For a very long pulse, the dislocations must "hover" at the CW nulls for a very long time before eventually traversing the rest of the trajectory, which in the CW limit they never get around to doing! So it is not so much that the trajectory is not there, just that it is not used.

In fact, we learnt in section 3.4 that the dislocation does not pass through the CW nulls, except in exceptional cases like the axial null of the piston radiator, where $\chi_o'' = 0$. Generally the effect of the term in braces in eqn.(3.22b), which we have so far ignored, will be to pull the trajectory away from the CW nulls by a small amount, which will decrease as $\sigma \to 0$ $(n \to \infty)$. So in the monochromatic limit the trajectory surface merges with the CW null lines, and the comments in the previous paragraph still apply.

It is interesting and unexpected to discover that for the piston radiator wavefunction and another unrelated wavefunction considered in a later chapter, the lines where $\partial M/\partial\omega = 0$ lie very close to the bottoms of valleys and the tops of ridges of M in space, and that the bottoms of valleys correspond to frequency minima and the tops of ridges to frequency maxima. This implies that the main effect of a small change in frequency is to locally shift the amplitude function rigidly in space in a direction perpendicular to its valleys, without a significant change in the actual value of the function otherwise. There appears to be no general reason why this should be so, therefore it must be a property of some restricted class of continuous wavefunctions of which the two wavefunctions we consider are members. It is not known just what this class is; for example it may be those wavefunctions containing a narrow angular spectrum of plane waves.

Another way of looking at the dislocation trajectories is to say that away from the CW nulls the zeros of the complex continuous wavefunction have moved out of real space into complex space. It can be shown (e.g. see Dennery & Krzywicki (1967), p98) that a complex

function of two (or, presumably, more) complex variables cannot have isolated singularities. Taking the function as $1/\psi$ implies that the simplest set of zeros of ψ as a function of two complex space variables is a complex surface which must extend to infinity. The valleys in the CW amplitude in real space are the "shadows" of these complex null surfaces (which may be closed loops). When the system is driven by a quasimonochromatic pulse there are times when the perturbation is sufficient to pull the complex null surface down into real space to produce a dislocation at some point in one of these "shadows".

3.7 Solution of the Full Perturbation Equations

If one compares fig.3.1e with fig.3.1b (or more easily with fig.3.2) one finds that χ_{KK} is very small along the lines $M_K = 0$. In fact, the zero contours of the two functions nearly coincide in places, with the notable exception of trajectory 1 for which χ_{KK} has no corresponding zero. There seems to be no general reason for this behaviour, but it means that as the frequency is varied through ω_o, the point representing ψ in the Argand plane moves approximately with constant speed round an arc of a circle centred on the origin. This suggests that our approximation of neglecting the term containing χ_{KK} should be better than we might expect.

Let us return to equations (3.22) and eliminate $f^{(2)}$ to give

$$\frac{f^{(1)}}{f} = -\pi n \frac{M^2 \chi_{KK}}{M_K M_{KK}}$$

which for our Gaussian pulse becomes

$$T + \chi_K = (\pi n)^2 \frac{M^2 \chi_{KK}}{M_K M_{KK}} \tag{3.24}$$

We must solve this equation simultaneously with one of equations (3.22), of which we will use (3.22a) in the form (3.23). Then the equation of the trajectory is

$$\frac{M^2 \chi_{KK}}{M_K M_{KK}} = \pm \sqrt{\frac{2M}{M_{KK}} + \frac{1}{(\pi n)^2}} \tag{3.25}$$

and the times are still given by (3.23) (or by (3.24)). Note that we cannot go back to our first approximation solution by letting $n \to \infty$, which gives

$$M_K = \pm \frac{M^2 \chi_{KK}}{\sqrt{2MM_{KK}}}, \quad \text{not } M_K = 0.$$

However, the right hand side is actually quite small, which is probably why our first approximation solution worked so well.

We see that now the dislocations in the head and tail of the pulse actually travel along separate trajectories, which we expect to be fairly close together. Let us examine the typical (we hope) behaviour of eqn.(3.25) by considering a section in the plane $Z = 0.5$ through trajectory 3 (indicated by the short vertical line on figs. 3.1b & e). Extracting values from figs.3.1, we plot in fig.3.6 the

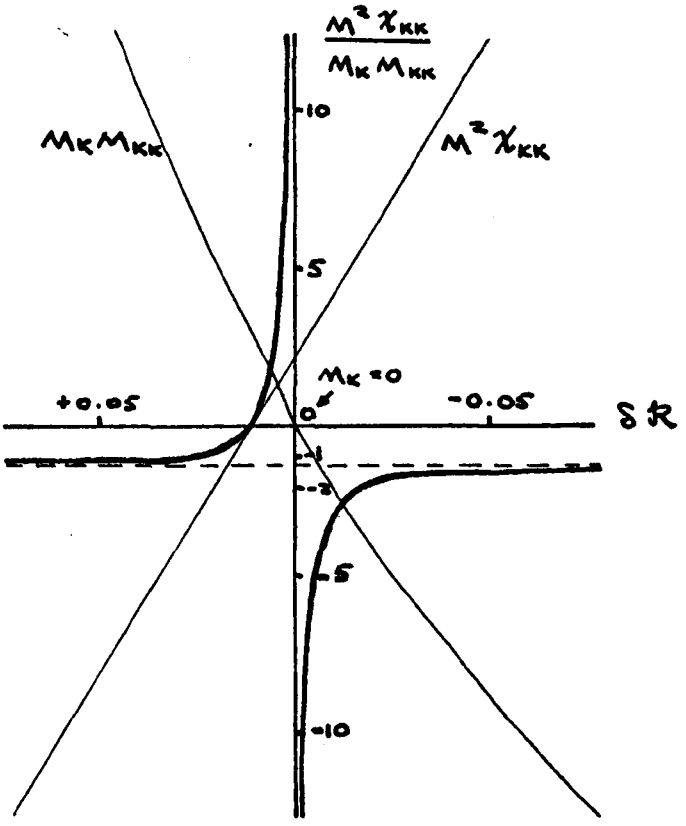


Figure 3.6a Typical Behaviour of Left Hand Side of Equation (3.25)

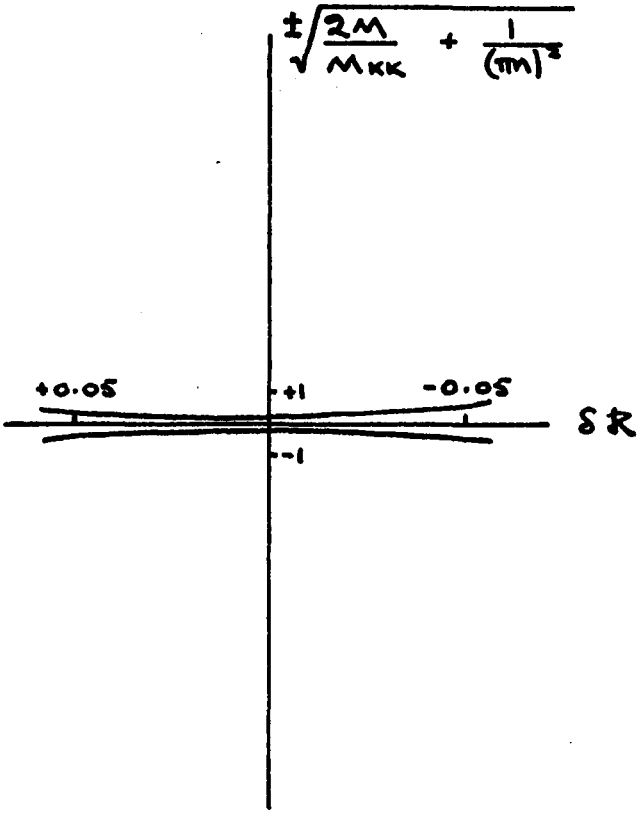


Figure 3.6b Typical Behaviour of Right Hand Side of Equation (3.25)

behaviour of the left and right hand sides of eqn.(3.25) in the neighbourhood of $M_K = 0$ (where $\delta\chi = 0$) for n = 3. The LHS is the quotient of two nearly linear terms with zeros a small distance apart, giving a curve like a rectangular hyperbola. On the same scale the RHS is a pair of very flat curves like parabolae. The trajectories are the intersections of the two graphs, which occur a small distance either side of where $\chi_{KK} = 0$, the positive-time solution being nearer to $M_K = 0$. As n is increased to ∞, the two branches of the RHS approach each other until they are slightly closer together than shown. As n is decreased to 0, the two branches separate (asymptotically as 1/n). The positive-time trajectory approaches where $M_K = 0$, moving very slowly The negative-time trajectory moves very rapidly away from where $M_K = 0$ (to infinity in this local model), jumps to the other side of where $M_K = 0$ and finally approaches $M_K = 0$. Thus the positive trajectory is stable to variations of pulse length, but the negative trajectory is highly unstable. Therefore we discard this unstable negative-time trajectory as being spurious, as we did previously without justification.

We can now proceed as before to compute the trajectories from the positive version of eqn.(3.25), and the times along the trajectories from eqn.(3.23). To represent eqn.(3.25) properly by its values at discrete points we must remove the infinite hyperbolic spike discontinuity by multiplying through by the denominator. We plot the zero contour defined by

$$M^2\chi_{KK} - M_K M_{KK}\sqrt{\frac{2M}{M_{KK}} + \frac{1}{(\pi n)^2}} = 0 \qquad (3.26)$$

which varies smoothly and continuously through the required zeros. Now there will be regions of space for which the square root in (3.26) does not exist, viz. the regions containing the spurious trajectories 1, 3 & 5. Therefore, the trajectory equation now automatically eliminates these spurious trajectories. Computationally we set (3.26) identically equal to zero where the square root does not exist, and then only plot the contours where the function crosses zero between consecutive lattice points.

The results of computations with the full perturbation equations are three trajectories (only) which are so close to trajectories 2, 4 & 6 of fig.3.2 that they are not worth showing. There are slight differences which can be seen if the two plots are superimposed, but they are within our "experimental error", and the full equations give

no obvious improvement. The most obvious difference is that trajectory 2 now stops short of the CW null (for purely technical reasons) although it should still pass through it exactly, because $\chi_{KK} = 0$ there. Similarly the times are virtually identical.

The increased computational difficulty of solving the full equations does not seem justified by the results: the main things we have achieved are to vindicate the first approximation solution (in practice, if not in theory: see the comments following eqn.(3.25)), and to give some theoretical justification for discarding the negative time solutions.

3.8 Continuous Wave Nulls

Since the CW nulls cause problems throughout this theory, we shall pause to consider them in a little more detail. In section 3.4 we avoided the singularities at CW nulls by an analytical trick, and deduced that if a dislocation passes through a CW null it does so at $f^{(1)}(\sigma\chi_0') = 0$ i.e. at the centre (in time) of a Gaussian-type pulse, as we saw for the axial null. Although the dislocations will only pass through the CW nulls in exceptional circumstances, they will normally pass near enough to experience the singularities. But we also deduced that M_2 and $f^{(2)}$ can only change sign at a CW null, and far from a null they must be both positive. Therefore the dislocation must be outside the inflexion of a Gaussian-type pulse right up to the null. So it appears that the dislocation jumps from outside the inflexion to the centre of the pulse at the null, and then back outside the inflexion again. This must be due to the singularities at the CW null, so what exactly are these singularities? We need to know what to expect, in order to have some hope of handling them computationally, because we cannot then perform our analytical trick to avoid the problem.

We shall deduce the likely form of the singularities which we expect to encounter as follows. As we vary the CW frequency ω, we expect the CW nulls to move around without much change in their local form. If the null passes through some point P at $\omega = \omega_0$, then the behaviour of ψ at P as a function of ω about ω_0 will be similar to the behaviour in space of ψ along the path of the null through P. We know that the dislocation trajectory is locally perpendicular to the path of the null, therefore we can also deduce the behaviour of ψ on the trajectory a small distance away from the null. The behaviour a small distance away along the null path is the same as at the null, but as for a slightly different ω. We sketch graphs of the behaviour in fig. 3.7. $\overline{\chi}'$ is a δ-function at the null, which broadens out into a Gaussian shape either side. But the actual δ-function will be lost by numerical interpolation, leaving two Gaussian-type humps either side of the null, with a dip in the middle. This is precisely what we see (half of) in fig.3.1d, and similar behaviour for $\overline{\chi}''$ in fig.3.1e. M' is a step function at the null which smooths off away from it, as verified by fig.3.1b, and M" is exactly as $\overline{\chi}'$, but the graph in fig. 3.1c is a bit more smoothed out because M" is derived from already-once-smoothed functions.

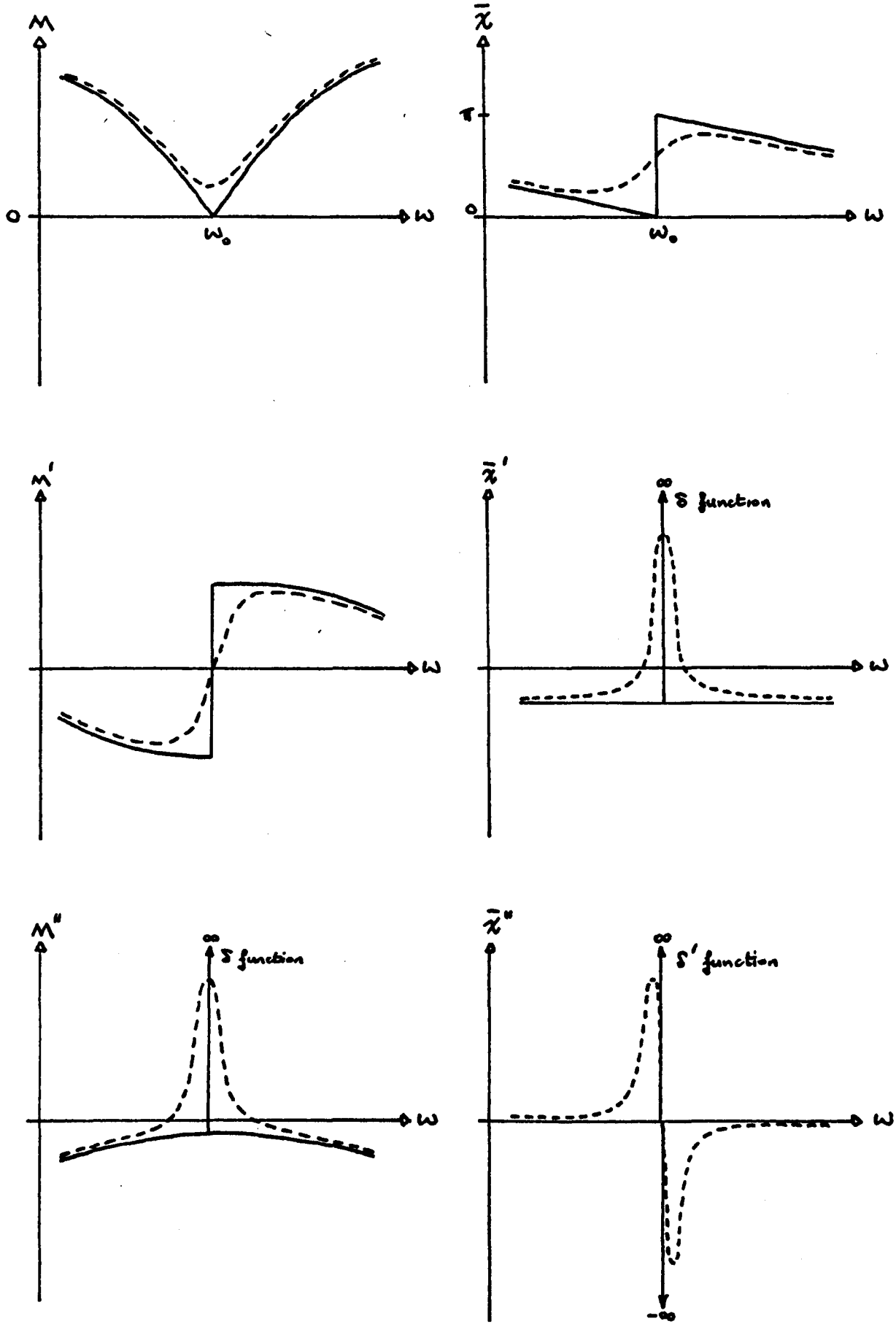


Figure 3.7 CW Null Singularities

The "centre" of the pulse is at $\sigma\chi_o' = 0$ i.e. $t = -\bar{\chi}_o'$, and we see that as we approach a CW null the "centre" of the pulse appears to occur at infinite time. This is obviously because the pulse amplitude is always very small at the CW null, so that the position of the pulse envelope is not really defined at all. Therefore, when we use the phase as normally defined, the dislocation is outside the inflexion. But if we change the definition of the phase to remove the singularity, the dislocation need no longer be outside the inflexion. We realise that the concept of the "position of the pulse" is somewhat nebulous, and depends on precisely how we define the phase. But although the position of the dislocation in the pulse is not really well defined, the actual time of the dislocation should be. At a CW null it is given by

$$t = -\bar{\chi}_o'$$

where $\bar{\chi}_o$ is analytic. Elsewhere it is given by

$$t = -\bar{\chi}_o' + \frac{1}{\sigma}\sqrt{\frac{2M}{\sigma^2 M_2} + 1}$$

Both terms on the right of this are singular at a CW null, and the singularities must cancel to give the same value as above. We should be able to evaluate this expression by taking the limit as we approach the null, or computationally by interpolating the whole right hand side through the null.

3.9 Appraisal of the Theory

Our results indicate that with this perturbation theory it is easy to find the spatial dislocation trajectories, but very difficult to find the times with any reliability. This is probably because the spatial trajectories are given to a very good approximation by just the very simple equation $\partial M/\partial \omega = 0$. This is independent of the pulse shape and length (although for too short a pulse the approximation would break down), while the time depends critically on the detailed shape and length of the pulse envelope. Our "experimental" pulse is probably too short for the theory to apply with any accuracy: it only contains 3 cycles of the carrier within its full width (of two standard deviations), and we saw in section 3.3 that it is not much longer than the critical length, below which additional solutions start appearing on the axis and in the far field of the piston radiator. It should be the case that for longer pulses the times would become more accurate, but this has not yet been examined.

A useful way to use the theory at present would be to predict (fairly reliably) all the spatial trajectories, and produce a rough estimate of the times. Then compute exact phase plots of areas along the trajectories at times around those predicted. This is not quite as "hit-or-miss" as the method of chapter 2, and we should at least be sure of not missing any dislocations completely. Let us apply this philosophy to the incipient dislocation which we have not yet properly investigated. This will allow us to check the predicted trajectory, and see if the times are at all reliable. At present all we know is that the top of the trajectory is correct. We display the results without comment in figs.3.8a & b (and we have also included them on figs.3.2b & 3.3c).

I can see no way to improve the present perturbation theory. Going to higher derivatives would be most unwieldy and probably not worth the trouble. The theory seems to work best where $\chi_o'' = 0$ identically, viz. on the axis and in the far field of the piston radiator. This is probably because elsewhere χ_o' and χ_o'' are treated quite differently, with no fundamental justification for doing so. This suggests that a perturbation theory which treats χ_o'' properly might give better results: such a method might be to expand the pulse

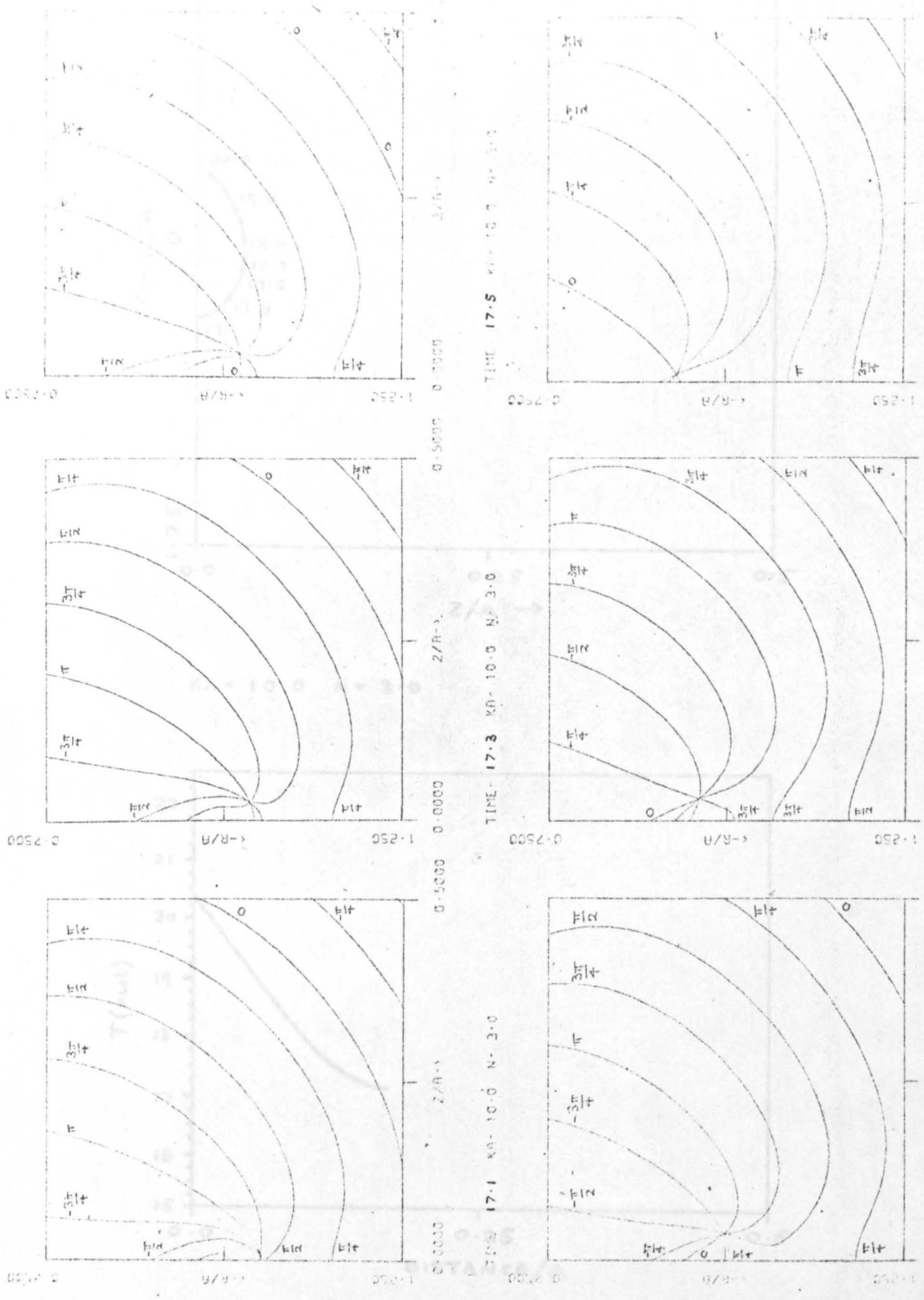
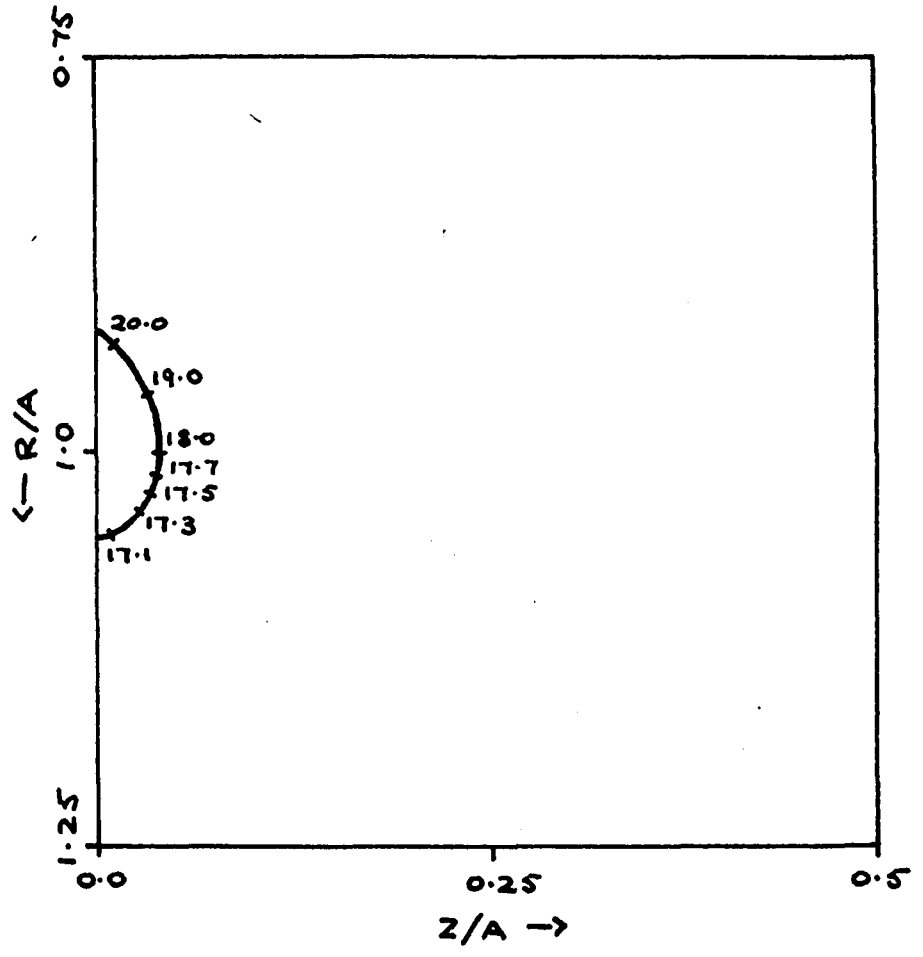


Figure 3.8a Equiphase Lines around Incipient Dislocation



KA = 10·0 N = 3·0

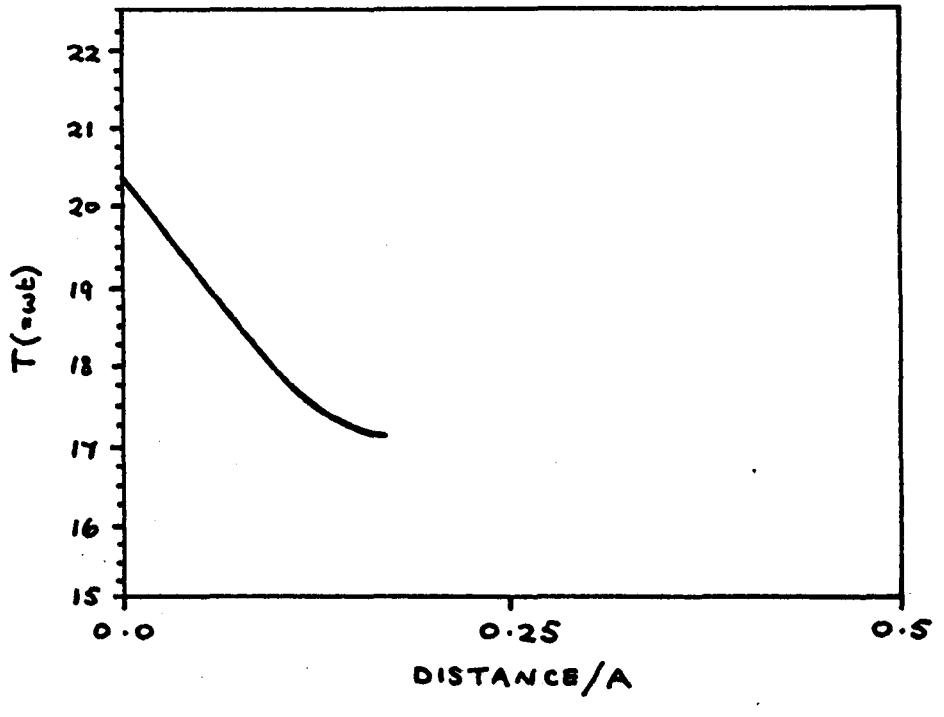


Figure 3.8b Trajectory and Times of Incipient Dislocation by exact computation

wavefunction in an asymptotic series by using a "steepest-descent" evaluation of the diffraction integral for a Gaussian-type pulse. This might give a different perturbation series in spectral-width, which may work better.

The most significant result of the present theory is that the trajectories are given to a very good approximation by the lines of frequency minima of CW amplitude, and the question remains: why? We shall attempt to shed some light on this in our concluding section.

3.10 Model Pulse with Very Simple Spectrum

We have seen that the behaviour in frequency space of the continuous wavefunction for a particular radiator is the key to the behaviour of the dislocations. Consider the pulse whose spectrum is just 2 discrete frequencies, viz.

$$\cos Wt\, e^{i\omega_0 t} = \tfrac{1}{2}\{e^{i(\omega_0 + W)t} + e^{i(\omega_0 - W)t}\}$$

If the continuous wavefunction at frequency ω is $\psi(\underline{r},\omega)e^{i\omega t}$ then the pulse wavefunction is

$$\Psi(\underline{r},t) = \tfrac{1}{2}e^{i\omega_0 t}\{\psi(\underline{r},\omega_0 + W)e^{iWt} + \psi(\underline{r},\omega_0 - W)e^{-iWt}\} \qquad (3.27)$$

Obviously the width of the spectrum is 2W, and we take this to be much less than ω_0. Let us assume that the phase of ψ is not stationary in frequency near ω_0, then using the notation $\psi_\pm \equiv \psi(\omega_0 \pm W)$ and $\psi_0 \equiv \psi(\omega_0)$, ψ_+ and ψ_- are two vectors in the complex plane either side of ψ_0. As t varies,

$$\tfrac{1}{2}\{\psi_+ e^{iWt} + \psi_- e^{-iWt}\}$$

traces out an ellipse, with semi-major axis of length $\tfrac{1}{2}\{M_+ + M_-\}$ at angle $\tfrac{1}{2}\{\chi_+ + \chi_-\}$, and semi-minor axis of length $\tfrac{1}{2}|M_+ - M_-|$, where

$$\psi = Me^{i\chi},$$

as in fig.3.9. The semi-major axis will not generally coincide with ψ_0.

This locus will only pass through the origin if $M_+ - M_- = 0$, i.e. the ellipse degenerates into a straight line. This requires all odd frequency derivatives of M to be zero, with no condition on the even derivatives. The CW amplitude must be stationary in frequency, but it can be a maximum, minimum or inflexion. When this condition is satisfied, the time is given by

$$2Wt = \pi - (\chi_+ - \chi_-) \quad (+ 2n\pi)$$

or $\quad t \simeq \pi/2W - \partial\chi/\partial\omega \quad (+n\pi/W)$ for small W.

The zeros of the transmitted pulse occurred at

$$t = \pi/2W \quad (+ n\pi/W),$$

so the diffraction appears to have shifted the pulse centre to $t = -\partial\chi/\partial\omega$, as we noticed before. In this model, the zeros in the original pulse occur where the CW amplitude is stationary in frequency, at a time $-\partial\chi/\partial\omega$ later.

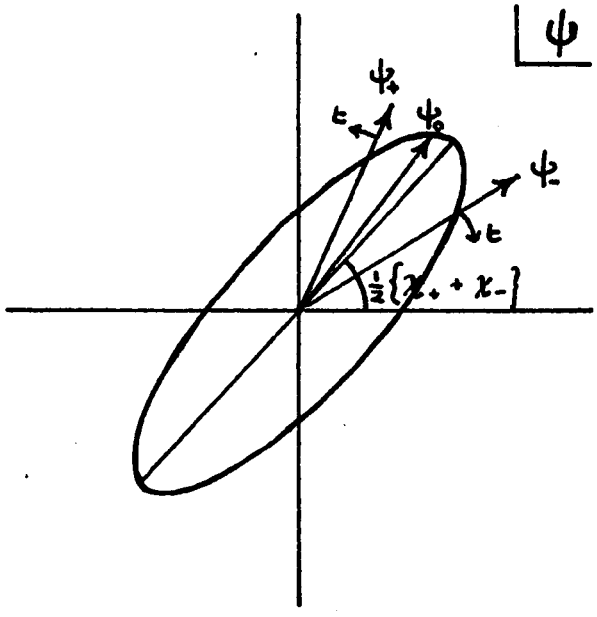


Figure 3.9 Locus of Term in Braces in Equation (3.27)

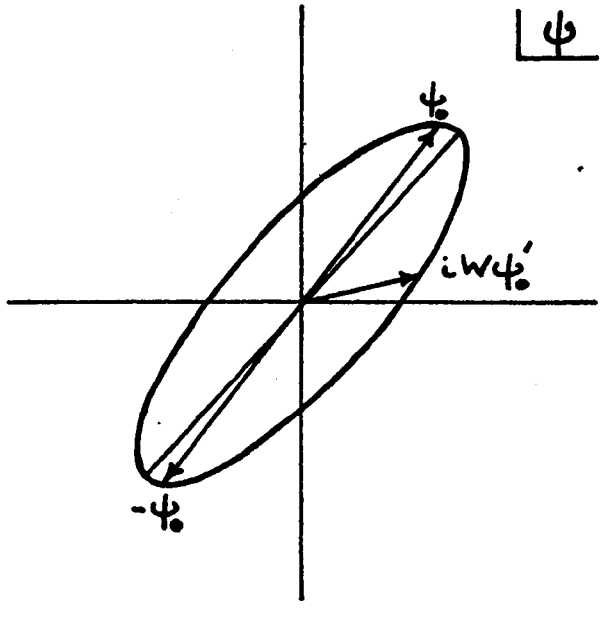


Figure 3.10 Locus of Term in Braces in Equation (3.29)

We would really like to consider a pulse which originally had no zeros, and the easiest way to do this is to add the centre frequency to give the pulse

$$(1 + \varepsilon\cos Wt)e^{i\omega_0 t} = e^{i\omega_0 t} + \tfrac{1}{2}\varepsilon\{e^{i(\omega_0+W)t} + e^{i(\omega_0-W)t}\}$$

whose width is still 2W. The pulse wavefunction becomes

$$\Psi(\underline{r},t) = e^{i\omega_0 t}\left[\psi_0 + \tfrac{1}{2}\varepsilon\{\psi_+ e^{iWt} + \psi_- e^{-iWt}\}\right] \tag{3.28}$$

For $W << \omega_0$ we can expand ψ as

$$\psi(\omega_0 \pm W) \simeq \psi(\omega_0) \pm W\psi'(\omega_0)$$

giving the condition for a dislocation as

$$\varepsilon\{\psi_0 \cos Wt + iW\psi_0' \sin Wt\} = -\psi_0 \tag{3.29}$$

The term in braces in (3.28) and (3.29) is exactly the same ellipse as we had before, and in this approximation passes exactly through ψ_0 at $Wt = n\pi$, and through $iW\psi_0'$ at $Wt = (n + \tfrac{1}{2})\pi$ (see fig.3.10). Therefore, we only have a solution of (3.29) in general if $\varepsilon = 1$, otherwise the locus of the left hand side will miss $-\psi_0$. But the transmitted pulse has second order zeros if $\varepsilon = 1$, so we shall avoid this case.

For $\varepsilon \neq 1$, the locus of the left hand side of (3.29) can only pass through $-\psi_0$ if it degenerates into a straight line along ψ_0. This happens if $iW\psi_0'$ is parallel to ψ_0. Now

$$iW\psi_0' = iW(M_0' + iM_0\chi_0')e^{i\chi_0}$$

$$= W(iM_0'e^{i\chi_0} - \chi_0'\psi_0)$$

is parallel to ψ_0 if $M_0' = 0$. So we see that $\partial M/\partial\omega = 0$ is a necessary condition on the amplitudes of the component frequencies of a pulse for them to be able to cancel out when the phases are right, to lowest order approximation, and this must remain so for continuous spectra.

For $M_0' = 0$ the locus of $\{*\}$ in (3.29) is a straight line slightly longer than ψ_0 (by $|W\psi_0'|$), and so if ε is not too small we should have a zero at some time. Although the geometrical argument is more illuminating, we can solve the whole problem analytically once we have made the approximation in eqn.(3.29), which becomes

$$M_0(1 + \varepsilon\cos Wt) + \varepsilon iW(M_0' + iM_0\chi_0')\sin Wt = 0$$

$\Rightarrow$ $M_0' = 0$ as we know

and $M_0(1 + \varepsilon\{\cos Wt - W\chi_0' \sin Wt\}) = 0$ if $M_0 \neq 0$. Then for small W

$$\cos W(t + \chi_0') = \pm \frac{1}{\varepsilon(1+(W\chi_0')^2)^{\frac{1}{2}}}$$

We must have $\varepsilon > 1 - \frac{1}{2}(W\chi_0')^2$ for a solution, and for small W, and ε near 1 we have $W(t + \chi_0') \simeq n\pi$, i.e. the dislocations are close to the maxima and minima of the transmitted pulse shifted in time by $-\chi_0'$ (again).

Returning to the exact expression (3.28) we know that for a dislocation to occur, the locus of {⇝} must be a straight line parallel to ψ_0, and if the phase of $\psi(\omega)$ is purely linear about ω_0, then this requires $M_+ = M_- = M_\pm$. Then the locus of

$$\tfrac{1}{2}\varepsilon\{\psi_+ e^{iWt} + \psi_- e^{-iWt}\}$$

is a straight line of length $\varepsilon M_\pm$ which must be $\geq M_0$ so that at some time the two terms of (3.28) can cancel, i.e. we require

$$M_\pm \geq M_0/\varepsilon .$$

For $\varepsilon < 1$ this implies M_0 is a frequency minimum. Therefore a pulse with no intrinsic zeros can only produce dislocations near to a frequency minimum of CW amplitude. But if $\varepsilon \geq 1$, M_0 may be a maximum or inflexion, so that if the pulse has intrinsic zeros then dislocations may appear wherever the CW amplitude is stationary in frequency.

Notice that for very small ε this model does not give any dislocations. In fact, Prof. J. F. Nye has shown that then the dislocations move in small ellipses centred on the CW nulls, rather than along the extended trajectories we have been considering.

APPENDIX

A3.1 Decomposition of a Complex Function

$$\psi(x) = u(x) + iv(x) = R(x)e^{i\chi(x)}$$

$$\Rightarrow \quad R = \pm\sqrt{u^2 + v^2}$$

$$\chi = \arctan(v/u) + n\pi$$

We require a convention to choose the sign of R and the value of n for χ. This choice can only change where $u(x) = v(x) = 0$, and these two equations define a subspace of x of codimension 2. If x is 2-dimensional, as in the standard theory of complex functions where x is a complex variable, then this "null set" is a point; if x is 3-dimensional it is a line, etc. It is possible to pass around the zero without passing through it because the space of x minus the null set is doubly connected around each null. Therefore, our choice of sign for R and value of n cannot change at all, so we choose $R \geqslant 0$ and $n = 0$.

Now if x is 1-dimensional, the null set is still a point, but now the space of x minus the null set consists of disconnected line segments, and we can choose the sign of R and value of n independently on the different segments. There are now two self-consistent possibilities: either $R \geqslant 0$ and $n = 0$, or $R > 0$, $n = 0$ and $R < 0$, $n = \pm 1$ on alternate segments, so that R and χ are both continuous and smooth.

CHAPTER 4
CAUSTICS AND CATASTROPHES

4.1 Introduction

In this chapter we begin our examination of the second type of singularity of wavefields, the caustics, which are really only singularities in the geometrical limit, as we mentioned in the prologue. The aim of this chapter is to introduce and motivate the detailed studies presented in subsequent chapters, and explain the general theory and methods to be used. Consequently there is little new material; rather a restatement of known results from the point of view required here.

In the geometrical limit "wave" energy travels in straight lines called rays (in a homogeneous medium, which is all that we shall consider). There is no diffraction because the wavelength is zero. At any point in the wavefield the energy arriving there is the sum of that carried by all the rays passing through the point. At most points there will be a small finite number of distinct rays, so that some small finite quantity of energy per unit time (intensity) arrives. But it may occur that all the rays leaving some small region of the radiator, of finite size, pass through a given point, of zero size. Then an infinite number of rays pass through that point, producing infinite intensity. Such an isolated point is called a point focus. If a perfect convex lens is held in sunlight, it will bring all the parallel rays from the sun which strike it to a point focus, as in fig.4.1. If a piece of paper is held in sunlight nothing happens. If the point focus is arranged to fall on the paper it will rapidly burn a hole in it, because all the energy which would have fallen on the dark disc, which is now in shadow, has fallen on the point focus. This ability of a focus to cause burning is the origin of the name caustic for the whole family of types of focus, of which a point focus is just a very special member, and is a striking illustration of the importance of caustics.

One sees that there are other types of focus than a point by examining closely the focal region of any real lens (the less 'perfect' the better), particularly if one looks at sections oblique to the incident rays. One then notices that what appears to be a bright point of light is really the point of a bright cusped conical surface, the tip of which is brighter than the rest. Focussing can also be produced

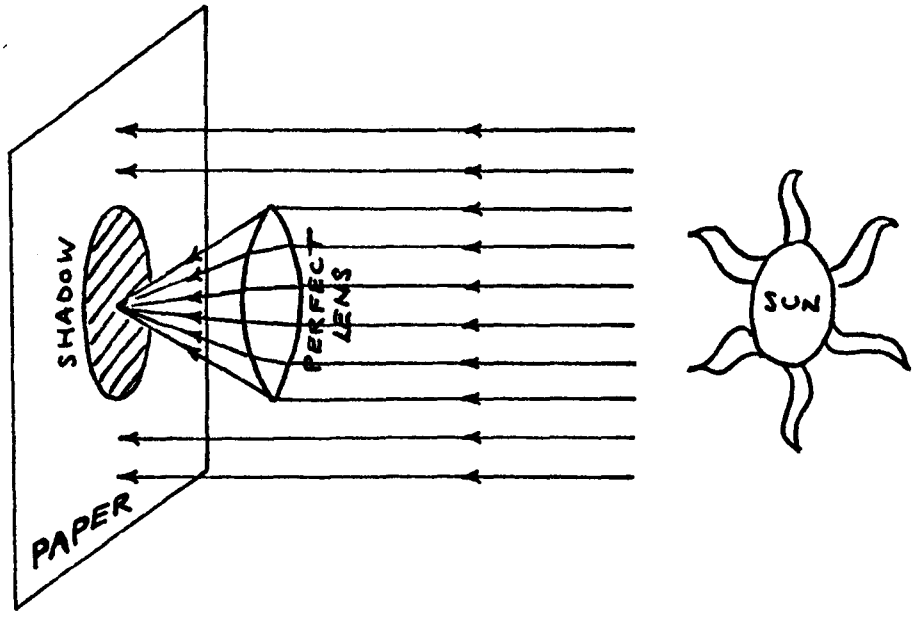
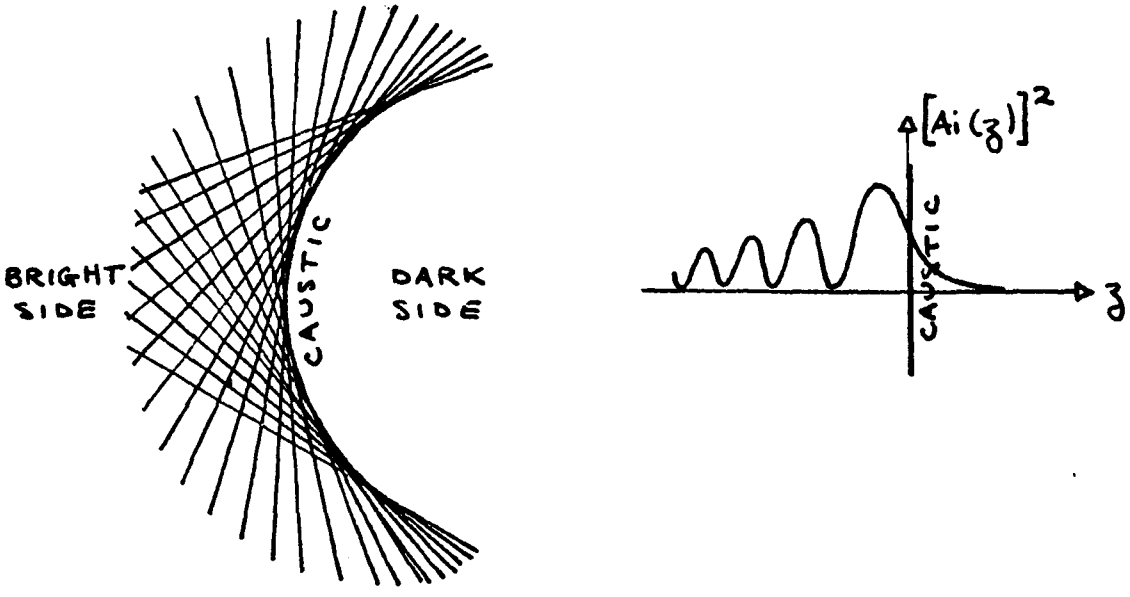


Figure 4.1 The Point Focus



(a) Ray Pattern (b) Diffraction Intensity

Figure 4.2 Line Caustic in Two-Dimensions

by reflection, and a good example of a caustic by reflection is the bright "heart-shaped" line seen on the surface of a cup of tea. Because of the cylindrical (i.e. translational) symmetry, the caustic must be a surface which is creased into a cuspidal line, and the surface of our beverage shows us a plane section through this. This cuspidal line of a caustic surface turns out to be generic, whereas the conical cusp produced by the lens is a more highly symmetrical non-generic case. For more examples of caustics produced by refraction through glass and water drops see Berry (1976), Nye (1978), by refraction and reflection by rippling water surfaces see Berry & Nye (1977), and for some beautiful caustics seen in the electron microscope see Liesegang (1953).

All the real caustics displayed in the above examples are produced by real waves with finite wavelengths, and consequently they are blurred out or disguised to some extent by diffraction effects, the success of the disguise being proportional to the wavelength. But the geometrical caustic for the problem still describes the most prominent features of the real diffraction pattern. The concept of rays is still useful if they are regarded as "carrying" the wave, so that instead of just adding the intensities of the rays reaching a point, one must add their actual wave amplitudes, taking regard of the phases. The rays are normal to the wavefronts at any point, and are also called wavenormals.

Generically, a bundle of rays do not all focus at the same point; rather, as one considers successive rays their point of focus moves, so that the rays converge along an envelope, rather than at a point. This produces a caustic surface, which is the "surface of centres" (i.e. the locus of the two centres of principal curvature, see e.g. HCV52) of the family of wavefronts, and generally has two sheets, which may interact. On one side, the bright side, of the caustic there are two different rays through any point; on the other side, the dark side, there are none (see fig.4.2a). Actually on the caustic there are two identical rays through any point: the two rays on the bright side have coalesced. On the caustic the two rays are exactly in phase, so one might expect the brightest part of the real diffraction pattern to lie exactly along the caustic. However, in reality, although the geometrical rays provide the main contribution to the wavefield, the rest of the wavefront also plays a small part, and this shifts the brightest region of the

diffraction pattern a little way onto the bright side of the geometrical caustic. It is this bright region that one actually "sees" as the caustic. There may also be other independent ray systems passing through the caustic, without affecting it appreciably. Therefore the criterion for a caustic surface is actually that the number of geometrical rays changes by 2 on crossing it.

4.2 Caustic Diffraction Patterns

The actual wavefunction near to a simple caustic surface was analysed long ago by Airy (1838), and the variation perpendicular to the caustic is locally the Airy integral function Ai(z) (see fig.4.2b, taken from Abramowitz & Stegun (1965)) where the geometrical caustic is at z = 0. This "Airy pattern" clearly shows the main maximum just inside this point, followed by a sequence of dark and light interference fringes of slowly decreasing visibility. Similarly the wavefunction near a cusp of a caustic was analysed by Pearcey (1946). The gross features of this are two superimposed Airy patterns, but we shall discuss this further later on.

The basic three-dimensional caustic is the smooth caustic surface, and we have seen that this surface can have a singularity in the form of a cuspidal line. Can it have other singularities? Can it have corners where two sheets meet at a finite angle, or is the angle restricted to $0^\bullet$ or $180^\bullet$? Can a cusp line itself have singularities, such as cusps or corners? How can caustic surfaces interact, and can they end? Obviously the number of globally different caustics is infinite, but we might hope to analyse a caustic into its elementary building blocks, if there are a finite number of these. If we can perform this local classification of caustics, we can then ask: does a local caustic structure carry a unique diffraction pattern, in any sense, and if so what is it? To answer these questions we must set up the mathematics of the problem. To do this we mainly follow Berry (1976), whose article provided the motivation for the study in the next few chapters, to which he alludes.

A continuous wavefunction (at time 0) can be written either exactly or approximately as a diffraction integral in the form:

$$\Psi(\underline{X}) = \int_S d^2\underline{x}\; a(\underline{x},\underline{X})\; e^{ik\Phi(\underline{x},\underline{X})} \qquad (4.1)$$

where the integral is over some surface S, which is essentially the wavefront, parametrised by $\underline{x}$, some distance away from the field point $\underline{X}$. $\Phi(\underline{x},\underline{X})$ is the distance from the source point $\underline{x}$ to the field point $\underline{X}$, and $a(\underline{x},\underline{X})$ is slowly varying. In chapter 2 we saw an exact example of this, and later we shall present an approximate example. At a particular point $\underline{X}$, the exponential will oscillate many times for a small change in $\underline{x}$, while $a(\underline{x},\underline{X})$ does not change appreciably, so that

the oscillations nearly completely cancel giving a very small contribution to $\Psi(\underline{X})$ from that neighbourhood of $\underline{x}$. But if Φ does not change much in the neighbourhood of some $\underline{x} = \underline{x}_0$, this will not be the case, and this neighbourhood of $\underline{x}$ will give a large contribution to $\Psi(\underline{X})$. Points where the phase Φ is stationary, i.e.

$$\nabla_{\underline{x}} \Phi(\underline{x},\underline{X}) = 0 \tag{4.2}$$

therefore give the main contributions to Ψ. About $\underline{x}_0$, Φ has the Taylor expansion

$$\Phi(\underline{x},\underline{X}) = \Phi(\underline{x}_0,\underline{X}) + \tfrac{1}{2}\{(\underline{x}-\underline{x}_0).\nabla_{\underline{x}}\}^2\Phi(\underline{x}_0,\underline{X}) + \ldots \tag{4.3}$$

as long as the stationary phase points (SPPs) $\underline{x}_0$ are well separated. Substituting this expansion to second order into (4.1) we can do the integral, and write $\Psi(\underline{X})$ as a sum of contributions by SPPs, by taking $a(\underline{x},\underline{X}) = a(\underline{x}_0,\underline{X})$ outside the integral.

But the condition (4.2) implies that $\underline{X}$ lies on the normal to the wavefront surface S at $\underline{x}$, which is the geometrical ray from $\underline{x}$. Therefore there is a one-to-one correspondence between geometrical rays and SPPs. A geometrical caustic or infinity of intensity occurs when two rays coalesce at a point. When two SPPs, i.e. two roots of (4.2), coalesce Ψ also becomes infinite <u>in the above approximation</u>. This is because Ψ contains the Hessian $H_{\underline{x}}\Phi$ in its denominator, and the vanishing of the Hessian is the condition for coincident roots of (4.2) (see eqns.(22) & (23) of Berry (1976)). We see that when two (or more) SPPs coalesce, the local expansion (4.3) of Φ breaks down. We need to know the possible local forms of $\Phi(\underline{x},\underline{X})$ about points where

$$\left.\begin{aligned} \nabla_{\underline{x}} \Phi(\underline{x},\underline{X}) &= 0 \\ H_{\underline{x}} \Phi(\underline{x},\underline{X}) &= 0 \end{aligned}\right\} \tag{4.4}$$

The corresponding loci of $\underline{X}$ are the local forms of the caustic, and from the local forms of Φ we can compute the local diffraction pattern around the caustic, so the caustic and its diffraction pattern are obviously in one-to-one correspondence.

4.3 Catastrophe Theory

The mathematics we need to use is called "Elementary Catastrophe Theory", in which the local forms of $\Phi(\underline{x},\underline{X})$ about a "catastrophe point", where (4.4) is satisfied, are classified by Thom's Theorem (Thom (1975), proved in full rigour by Mather (1972)). In this theory $\underline{X}$ are called "control variables" because they are under our control: the point at which we choose to examine the wavefunction. $\underline{x}$ are called "state variables" because they tell us the state of the system and are not directly under our control: they label the rays passing through $\underline{X}$. $\Phi(\underline{x},\underline{X})$ is called the potential function because its "minimization" gives the "equilibrium condition" for the rays. We shall describe Thom's theorem in sufficient detail to apply it to our problem, but without any attempt at full mathematical rigour (for which see Wassermann (1974), Bröcker (1975)).

Only smooth C^∞ manifolds and maps are considered. Let $\underline{X} \in$ control space $C = \mathbb{R}^k$ $(k \leq 6)$, $\underline{x} \in$ state space S = smooth n-manifold, and F be the space of smooth maps $\Phi : C\times S \to \mathbb{R}$ equipped with the Whitney C^∞ topology. Then Thom's theorem states that there exists an open dense subspace $F_0 \subseteq F$ such that if $\Phi \in F_0$:

1) the solutions of $\nabla_{\underline{x}}\, \Phi(\underline{x},\underline{X}) = 0$ form a smooth submanifold $M^k \subseteq C\times S$

2) for such a Φ, any singularity of the projection

$$\chi : M^k \to C$$

is equivalent to one of the standard list of elementary catastrophes

3) χ is then stable, in the sense that it has a neighbourhood in F_0 of equivalent maps.

This notion of equivalence is very important. The functions $f : M \to Q$ and $f' : M' \to Q'$ are equivalent if there exist diffeomorphisms $h : M \to M'$ and $k : Q \to Q'$ such that the diagram

$$\begin{array}{ccc} M & \xrightarrow{f} & Q \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle k} \\ M' & \xrightarrow[f']{} & Q' \end{array}$$

commutes. This means that for a given starting point in M we end up at the same point in Q' if we apply h then f', or if we apply f then k. Therefore we can start at any point, go right round the loop either way

and end up back at the same point, because $f^{-1}k^{-1}f'h = 1$. Loosely we say that f and f' are <u>diffeomorphic</u> if they are equivalent as above.

We are here only interested in caustics in real three-dimensional space, which have no other "internal" control parameters, and therefore we shall restrict ourselves to the catastrophes which can arise in "codimension" $k \leqslant 3$. Following Berry (1976), but using $\underline{W}$ for our standard control variables (because computers do not understand Greek), the list of standard elementary catastrophes is:

codim	name		standard potential function $\phi(\underline{x},\underline{W})$
1	fold	cuspoids	$\frac{x^3}{3} + W_1 x$
2	cusp		$\frac{x^4}{4} + W_1\frac{x^2}{2} + W_2 x$
3	swallowtail		$\frac{x^5}{5} + W_3\frac{x^3}{3} + W_2\frac{x^2}{2} + W_1 x$
3	elliptic umbilic	umbilics	$x^3 - 3xy^2 - W_3(x^2+y^2) - W_1 x - W_2 y$
3	hyperbolic umbilic		$x^3 + y^3 + W_3 xy - W_1 x - W_2 y$

Here x and y are "essential" state variables. The catastrophes which may occur in a given codimension are independent of n, the dimensionality of the state space: any state variables which exist in addition to those shown above can only occur in the potential function in an inessential way, typically quadratically, so that they produce no degenerate critical points (see Poston & Stewart (1976)). Catastrophes with one essential state variable are called cuspoids, those with two are called umbilics (because umbilic caustics are produced by the neighbourhood of umbilic points on a wavefront, where the two principal curvatures are equal).

There is an important distinction between cuspoid and umbilic catastrophes, mentioned by Woodcock & Poston (1974). The projection χ is (locally) a map $\mathbb{R}^k \to \mathbb{R}^k$, but the definition of catastrophes makes them distinct from singularities of maps $\mathbb{R}^k \to \mathbb{R}^k$ stable with respect to perturbations of the map, and neither class contains the other. The intersection of the two classes consists of the cuspoids, which are stable to perturbations of Φ and χ. The umbilics, however, are stable only to perturbations of Φ, and are not map stable.

Let us consider the meaning of Thom's theorem as applied to caustics. Any realistic Φ will satisfy the smoothness conditions of the theorem (or can be approximated arbitrarily closely by a function that does). For illustration let us consider $k = 2$, $n = 1$ so that $C \times S$ is a visualizable three-dimensional space, as for the cusp catastrophe shown in fig.4.3. Point (1) above tells us that generically the ray condition $\nabla_{\underline{x}} \Phi(\underline{x},\underline{X}) = 0$ restricts us to a smooth two-dimensional surface in $C \times S$, called the equilibrium or catastrophe manifold. If we choose a point in the C plane, and draw a line parallel to the S axis, every intersection of this line with the catastrophe manifold gives a geometrical ray. So the number of sheets of the catastrophe manifold "above" a point of C tells us the number of rays reaching that point. Since the manifold is smooth the only way that two sheets can link up is along a smooth fold of the surface. We could draw a smooth continuous line along this fold, where the surface is perpendicular to C. Then it is precisely along this line that two rays will coalesce. Woodcock & Poston (1974) have computed pictures of all the catastrophe manifolds we are interested in, and these clearly show the fold lines. The projection of this fold line into the control space is called the bifurcation set, because it bifurcates the control space (locally) into two regions in which the number of equilibrium points (rays) differs by two. We see that this is precisely the geometrical caustic.

The fold line is also the line along which the projection χ is singular, because it corresponds to a distinct edge of the manifold as far as the projection is concerned (strictly χ is singular because we cannot use the implicit function theorem to solve for $\underline{x}$ as a function of $\underline{X}$ along the fold line i.e. χ^{-1} does not exist, e.g.see Flett (1966), pp400-5). Point (2) tells us that in the neighbourhood of any point on the fold line taken as origin, Φ must be locally diffeomorphic to one of the standard potential functions ϕ. This means that there exists a smooth invertible change of coordinates which will cast our Φ into a form diffeomorphic to a standard ϕ, plus higher order terms which we can ignore since we are only interested in local behaviour. We are justified in doing this because the determinacy of our original function will be the same as our standard function, about the origin (see Poston & Stewart (1976)).

What Thom's theorem does not tell us, is what the diffeomorphisms to apply are, or which standard potential function we are trying to map

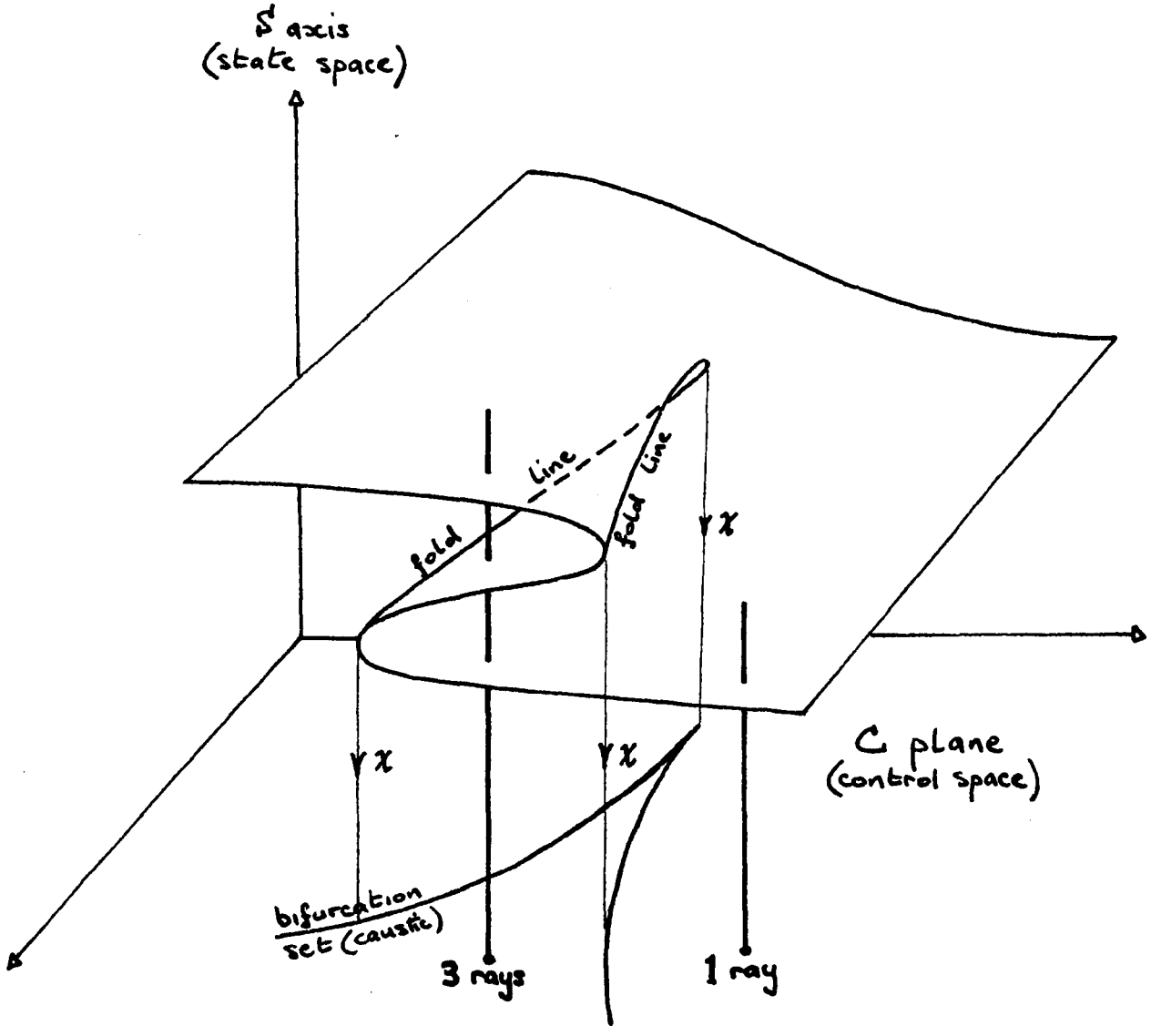


Figure 4.3 Cusp Catastrophe Manifold

onto, although often one has some physical reason for expecting a particular catastrophe. One can, in principle, find the caustic, and then at any point find the codimension, which is one less than the number of coalescing rays (see Berry (1976)). One can then expand Φ about the fold point in a Taylor series and compare with the standard forms for that codimension. We have, in effect, found a very local approximation to the required local diffeomorphism. We shall illustrate this in the next chapter. The caustic that we find will be diffeomorphic to the bifurcation set of the appropriate catastrophe.

Point (3) tells us that these caustics are structurally stable, in the sense that a small perturbation of Φ will not change the diffeotype of the caustic. This is crucial to the whole theory. It means that similar (i.e. diffeomorphic) caustics are produced by a variety of similar potential functions. Therefore, one has a good chance of observing one of these similar caustics in nature, because its production does not depend on anything special having occurred. Although the number of actual local caustic types is uncountably infinite, there are only 5 local caustic diffeotypes (in codimension 3), and it is precisely this fact which makes their study feasible.

The map stability of the cuspoids is additional to their structural stability. It means that if one changes the angle of the projection in the space C×S by a small amount, instead of projecting "vertically down" into C, then one produces a diffeomorphic bifurcation set. This can be clearly seen in the rotated stereoscopic pair pictures in Woodcock & Poston (1974). Small tilts of the cusp produce another cusp. Small tilts of 3-dimensional sections through the swallowtail manifold produce different 2-dimensional sections of the swallowtail bifurcation set, which are not always diffeomorphic, but if one could tilt the whole manifold in 4-dimensions it would produce a diffeomorph of the whole 3-dimensional bifurcation set. Similarly for the higher cuspoids.

This brings us to the point that although every bifurcation set is structurally stable, not every section through it is. The actual catastrophe point is at $\underline{W} = 0$ (in the standard forms), called the organizing centre, where the potential function reduces to its "germ", and the rest of the catastrophe constitutes an "unfolding" of the germ. Any section through the organizing centre is unstable, because a generic perturbation will move the organizing centre out of that section. Such

a section is non-generic (one's chances of choosing one at random are nil): any generic section is structurally stable. A generic section can only contain subordinate catastrophes, of lower codimension (e.g. see Connor (1976)), e.g. generic 2-dimensional sections through any of the 3-dimensional bifurcation sets can only contain cusps and folds, which one cannot necessarily attribute to the higher catastrophe. If, in 2-dimensions, one sees anything other than cusps and folds, one knows that one is seeing a singular section of a higher dimensional bifurcation set (which one can often identify, especially with the help of its canonical diffraction pattern). This must be due to some special circumstance obtaining, such as an unsuspected symmetry, a fact made good use of by Berry (1975).

4.4 The Canonical Catastrophe Diffraction Integrals

Having discussed the necessary mathematics, we can now answer the questions posed in section 4.2. The possible generic local forms of a caustic in three-dimensions are diffeomorphs of our 5 standard bifurcation sets (which are sketched in Berry (1976) and which we shall consider in detail one by one). Each of these local forms carries a specific type of diffraction pattern. If a particular catastrophe occurs at $(\underline{x}_0, \underline{X}_0)$ then very close to $\underline{X}_0$ the wavefunction, from (4.1), is approximately

$$\Psi(\underline{X}) \simeq a(\underline{x}_0, \underline{X}_0) \int_S d^2\underline{x}\, e^{ik\Phi(\underline{x},\underline{X})}$$

where $\Phi(\underline{x},\underline{X})$ is diffeomorphic to the standard potential function. If we define a set of canonical diffraction integrals

$$\psi(\underline{W}) = \left(\frac{1}{2\pi}\right)^{m/2} \int d^m\underline{x}\, e^{i\phi(\underline{x},\underline{W})} \tag{4.5}$$

where m = 1 for cuspoids and 2 for umbilics, then the diffraction pattern $\Psi(\underline{X})$ is diffeomorphic to $\psi(\underline{W})$. We can compute these canonical diffraction patterns once and for all, and then use them much like a set of logarithm tables. (The reason for the factor $(1/2\pi)^{m/2}$ in (4.5) is so that if inessential state variables were added to ϕ in the form $x^2/2$, they would not change the amplitude of $\psi(\underline{W})$.)

We may still wonder whether these canonical integrals have any quantitative use further away from a caustic. So far we have only used them in a "transitional approximation" very close to a caustic, by expanding the phase about the caustic, and setting $a(\underline{x},\underline{X})$ equal to the value on the caustic. Away from the caustic this breaks down, because the value of $a(\underline{x},\underline{X})$ changes, and terms of higher order than those retained in the phase expansion become significant. But here the simple stationary phase method works. Hence across a caustic we must make three different approximations. But these different approximations do not match up smoothly and hence to evaluate a wavefunction continuously through a caustic by these methods is very unsatisfactory.

However, we can make a "uniform approximation" to the wavefunction, valid uniformly through the caustic, by using the canonical integrals (4.5) as the "comparison integrals". The uniform approximation matches up smoothly with the stationary phase approximation sufficiently far from the caustic, and the changeover point is not at all critical. This is eminently satisfactory for computational work, and we shall use this

method to handle the coalescence of two complex SPPs in the swallowtail integral. The crux of the method is to find a mapping from $\Phi(\underline{x},\underline{X})$ to $\phi(\underline{x}',\underline{X})$ which is one-to-one, which is achieved by ensuring that the stationary points of Φ map onto those of ϕ. The details are discussed in section 5 of Berry (1976) and by Connor (1976): the result is an expression for $\Psi(\underline{X})$ in terms of $\psi(\underline{W})$ and its derivatives. Therefore a knowledge of the canonical integrals and their derivatives allows one to evaluate a wavefunction, defined by a diffraction integral, in a large region around a caustic, with very good accuracy. I have not actually computed any derivatives of the three-dimensional canonical integrals, but I believe that the numerical methods to be presented could easily be modified to do so.

The simplest canonical integral is that for the fold, which we will call

$$F(W_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\ e^{i(x^3/3 + W_1 x)} = \sqrt{2\pi}\ \mathrm{Ai}(W_1)$$

(see Abramowitz & Stegun). We see that the canonical catastrophe integral for the fold is precisely the Airy integral function, introduced by Airy to describe the wavefunction locally across a caustic surface. In the three-dimensional $\underline{W}$ space the caustic is the plane $W_1 = 0$. The Airy function has been tabulated and plotted long ago, so we need not discuss it further.

The next canonical integral is that for the cusp, which we will call

$$C(W_1,W_2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\ e^{i(x^4/4 + W_1 x^2/2 + W_2 x)}$$

This is essentially the integral used by Pearcey (1946) as a local approximation to the field near a cusped caustic. We shall discuss it in detail in the next chapter.

We see that the two simplest canonical integrals have been known for some time. Catastrophe theory has extended the set of such integrals and put their derivation on a rigorous theoretical basis. Of the higher canonical integrals, only the sections $W_3 = 0$ of the elliptic and hyperbolic umbilics have previously been computed, by Trinkaus & Drepper (1977), who compare them with the cusp pattern. In the following chapters we shall plot contours of amplitude and phase of the canonical

integrals for the cusp, swallowtail and elliptic umbilic (the hyperbolic umbilic has yet to be done). The cusp pattern is two-dimensional, so there is no problem, but the others are three-dimensional, and we can only plot two-dimensional sections. Since W_3 lies along the main axis of the bifurcation set for all the three-dimensional catastrophes, we shall plot sections W_3 = constant through them. There is no reason why other sections should not be plotted (in fact we shall also plot $W_2 = 0$ for the elliptic umbilic), but as an initial study the W_3 sections were expected to be the best choice.

4.5 Integration Method

We shall consider here the one-dimensional integral of the form

$$I = \int_{-\infty}^{\infty} dx\ e^{i\phi(x)}$$

where $\phi(x)$ is a polynomial in x, whose coefficients depend on the control variables $\underline{W}$ linearly. At both ends of the range of integration $\phi(x)$ diverges, so the integrand oscillates infinitely rapidly with a constant amplitude of 1. This is the main problem in numerical integration of these integrals. By rotating the contour of integration through $\pm\pi/2n$ in the complex x plane, where n is the order of the polynomial ϕ, one can turn the integrand into an oscillatory function whose amplitude asymptotically decays exponentially for $n > 1$. (If n is odd the two half integrals from 0 to $\pm\infty$ must be treated separately). This shows that the integrals converge (and also allows one to evaluate them in terms of Γ-functions when ϕ reduces to its monomial germ at the organizing centre $\underline{W} = 0$, which provides a useful check on the numerical results). However, under this transformation, in certain regions of control space, the amplitude of the integrand becomes extremely large before the exponential decay takes over. Since the values of the integrals are of order unity, this method is no use numerically.

A standard method for handling infinite range integrals numerically is to map the infinite range onto a finite range of some new integration variable. But this would merely compress the infinitely rapid oscillations still further, and it is these oscillations which are essentially the problem, not the infinite range, so this method is not applicable. Maslin (1976) has used a double convergent series expansion in $\underline{W}$ for the cusp diffraction integral. However, for the three-dimensional catastrophes this would involve a triple series with complicated coefficients, which would only converge rapidly near the origin. It would probably need to be supplemented by an asymptotic expansion further from the origin. Such a method was rejected as being not sufficiently general. It would probably require arbitrary matching of different solutions, and there is no reason to expect it to be any more efficient than the direct integration method used here. It might be possible to estimate the errors more accurately, although for just plotting contours high accuracy is not necessary; but a large number of data points, and therefore high efficiency, is necessary.

The method chosen is based on the ubiquitous stationary phase method, but rendered exact (in principle) and automated. The form of output required has a large bearing on the choice of integration method. To plot contours of a function defined by a complicated integral, the best method is to evaluate the function on a lattice of closely spaced points. This fact can be capitalized upon by stepping from one point to a neighbouring point, and using iteration starting from the value at the previous point. The main contribution to I will come from the neighbourhood of points where $\partial\phi/\partial x = 0$, and therefore these regions should be integrated exactly by quadrature. The contributions from the highly oscillatory tails will be small, and we can expand these contributions in asymptotic series. The integral is stable to changes in the cutoff, at which we change from quadrature to series solution, (as it must be for the method to be reliable) so that we can vary this parameter to optimise the efficiency.

The method was used in its basic non-optimising form by Berry (1975) to compute the intensity of the cusp diffraction pattern shown in his fig.4, but was found to be highly prone to "spurious convergence" when using Simpson's rule quadrature. This is a common problem in quadrature of oscillatory integrals, and arises from the ambiguity in specifying a continuous function by its value at discrete points. The integration routine can quite easily find two successive approximations which agree to some required accuracy, but are actually both wrong, which is very hard to check for automatically. Fortunately, this error normally shows up quite clearly on contour plots, and with the current optimising algorithm, and a sophisticated quadrature routine, when the various parameters are chosen correctly, spurious convergence is very rare, and the graphs displayed in subsequent chapters are "straight off the computer".

The asymptotic expansion of this type of integral is discussed by Dingle (1973). Consider

$$I = \int_A^B dx\, e^{i\phi(x)} = \int_{\phi(A)}^{\phi(B)} \frac{d\phi}{\phi_1}\, e^{i\phi}$$

(using the notation $\phi_n(x) = \frac{d^n\phi(x)}{dx^n}$), where we have changed the variable of integration to ϕ. If there are no stationary points of ϕ in the range $[A,B]$ then the second integrand is regular everywhere. Integrating

I by parts m times gives (Dingle, p114)

$$I = \left[-ie^{i\phi}\sum_{r=0}^{m-1}\left(i\frac{d}{d\phi}\right)^{r}\frac{1}{\phi_1}\right]_{\phi(A)}^{\phi(B)} + \int_{\phi(A)}^{\phi(B)} d\phi\left\{\left(i\frac{d}{d\phi}\right)^{m}\frac{1}{\phi_1}\right\}e^{i\phi}$$

We shall use this formula either with $B = \pm\infty$, or with $A = 0$ (in which case ϕ will include inverse powers of x, as in the elliptic umbilic) and the contribution to the series of these end points will be zero. Then the above integration process is precisely equivalent to introducing in the integrand the Taylor series for $1/\phi_1$ as a function of ϕ about the finite end point C, say (see Dingle pp111-2). If $\phi(x)$ is an nth order polynomial, then $1/\phi_1$ has (n-1) poles in the finite complex ϕ plane, and its Taylor series only converges within the circle centred on $\phi(C)$ which touches the nearest of these. In the case $B = \pm\infty$, the series is integrated term by term from $\phi(A)$ to $\pm\infty$, which must extend outside the circle of convergence. Therefore the series for I is asymptotic, rather than convergent. In the case $A = 0$, the range of integration $[0,\beta]$ is finite (and very small in practice), so that β can be chosen small enough that $[0,\beta]$ lies within the circle of convergence and the series for I is not necessarily divergent (but see appendix A7.1).

We shall concentrate here on the asymptotic series, and we must consider its convergence. We expect that if ϕ_1 is made large enough the series can be made to converge as rapidly as required (before it begins to diverge). The size of the remainder integral R_m is given by

$$|R_m| \leq \left|\int_{\phi(A)}^{\pm\infty} d\phi \left(\frac{d}{d\phi}\right)^{m}\frac{1}{\phi_1}\right| = \left|\left(\frac{d}{d\phi}\right)^{m-1}\frac{1}{\phi_1}\right|_{\phi(A)}$$

if $\left(\frac{d}{d\phi}\right)^{m-1}\frac{1}{\phi_1}$ is monotonic in the range $[\phi(A),\pm\infty]$. Then the truncation error will be less than the magnitude of the last term included. Similar results can be derived in terms of the term after the last term included (see Erdelyi (1956)), but computationally an estimate of the error in terms of the quantities already calculated is more useful. In fact, the above result is likely to be a gross overestimate of the error, since it completely ignores the oscillatory nature of the true integrand, and the monotonicity condition is probably overrestrictive. It is also difficult to test for. Sufficient conditions for monotonicity can be found in terms of the complex roots α_j of $\phi_1(x) = 0$ as follows:

$m = 1 \qquad x \geqslant \mathrm{Re}\ (\alpha_j) \qquad\qquad\qquad$ for all j

$m = 2 \qquad x \geqslant \mathrm{Re}\ (\alpha_j) + |\mathrm{Im}\ (\alpha_j)| \qquad$ for all j

$m = 3 \qquad x \geqslant \mathrm{Re}\ (\alpha_j) + \sqrt{3}\ |\mathrm{Im}\ (\alpha_j)| \qquad$ for all j etc.

When all the roots are real these conditions must be met anyway to avoid stationary points. Otherwise, numerical tests show that they are over-restrictive, and computing the complex roots of $\phi_1(x) = 0$ is not convenient in general, anyway. However, these results do prove that if the cutoff A is made large enough that the series is converging well, then the truncation error will be less than the last term included.

Some idea of the behaviour of the series can be derived by assuming the cutoff is large enough that $\phi(x) \sim x^n/n$. Then

$$\frac{1}{\phi_1} \sim x^{1-n}$$

$$\left(\frac{1}{\phi_1}\frac{d}{dx}\right)\frac{1}{\phi_1} \sim (1-n)\, x^{1-2n}$$

$$\left(\frac{1}{\phi_1}\frac{d}{dx}\right)^2\frac{1}{\phi_1} \sim (1-n)(1-2n)\, x^{1-3n}$$

and the monotonicity condition is clearly satisfied. The method adopted in practice is to take the first three terms of the series, to test that they are converging and that the last term is less than the maximum permitted error. This seems to be quite satisfactory provided the convergence conditions are properly chosen: the initial trial parameters can be deduced form the above formulae.

We define

$$G(x) = e^{i\phi(x)} \left\{ i\left(\frac{1}{\phi_1(x)}\right) + \left(\frac{\phi_2(x)}{\phi_1{}^3(x)}\right) + i\left(\frac{\phi_1(x)\phi_3(x) - 3\phi_2{}^2(x)}{\phi_1{}^5(x)}\right) \right\} \qquad (4.6)$$

$\to 0$ as $x \to \pm\infty$. Then

$$\int_A^{\pm\infty} dx\, e^{i\phi(x)} = \left[-G(x)\right]_A^{\pm\infty} = G(A)$$

and

$$I = \int_{-\infty}^{\infty} dx\, e^{i\phi(x)} = \left(\left(\int_{-\infty}^{B} + \int_B^A + \int_A^{\infty}\right) dx\, e^{i\phi(x)}\right.$$

$$= G(A) - G(B) + \int_B^A dx\, e^{i\phi(x)} \qquad (4.7)$$

We attempt to scan the control plane in such a way that the SPPs, and hence the cutoffs A and B, separate. We start with the cutoffs

outside the outermost zeros of $\phi_1(x)$, where the magnitude of $\phi_1(x)$ is large enough to make the first term of G(x) sufficiently small. As the zeros of $\phi_1(x)$ move out, the magnitude of $\phi_1(x)$ at the cutoff falls, and hence the cutoff must move out. Therefore, the condition that $|\phi_1(\text{cutoff})|$ is large enough ensures that the cutoffs remain outside all stationary points, unless the bifurcation set is crossed in such a way that a new pair of zeros of $\phi_1(x)$ appear outside the cutoffs. This situation must be detected, and the cutoffs made to jump outside the outermost new zero. Having found the approximate position of the cutoffs, we then push them out until G(x) converges adequately. Once we have found the correct values of A and B, we compute G(A) and G(B), and evaluate the integral over the stationary phase points by quadrature. Since the cutoffs only move outwards as we compute successive data points, we can use the previous cutoffs as a first estimate at the next point. This is the iteration mentioned above. We only have to actually find the SPPs at the start of the scan. From then on the cutoffs are determined by successively evaluating the terms in G(x), which we will need anyway, and only stepping the cutoffs out when necessary for convergence. In this way they are found and optimised efficiently. Of course, if the cutoffs are further out than necessary G(x) converges very rapidly, but the range of the numerical integration is wider than it need be. This increases the time needed to perform the quadrature, and because of the nature of the integrand greatly increases the risk of spurious convergence. Hence the need to optimise the cutoffs very carefully!

The precise details of the integration method depend on the integral, and we shall discuss them in a little more detail in appendices to the appropriate chapters. The techniques have also evolved as more experience has been gained. The method was initially tested on the fold integral, and the results compared well with tabulated values of the Airy function. The cusp integral also provides a good test of the method, as we shall see in the next chapter.

CHAPTER 5

THE CUSP DIFFRACTION PATTERN

5.1 The Canonical Integral and Caustic

In this chapter we primarily study the integral

$$C(W_1,W_2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\ e^{i\phi(x,W_1,W_2)} \tag{5.1}$$

where $\phi(x,W_1,W_2) = x^4/4 + W_1x^2/2 + W_2x.$ (5.2)

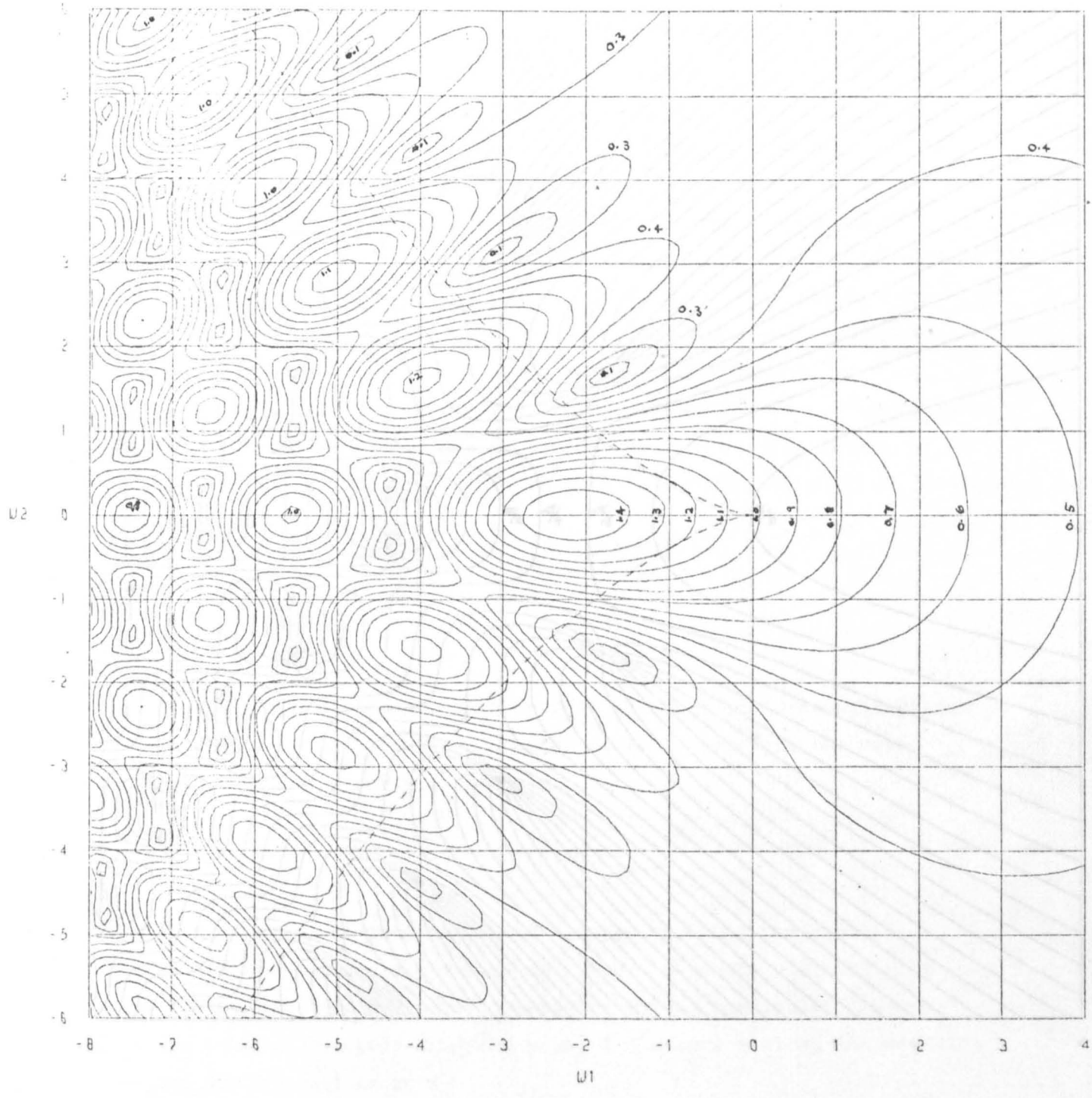
In figs.5.1 & 5.2 we show contour plots of the amplitude and phase of the complex function $C(W_1,W_2)$. The details of the numerical integration and methods of plotting phase contours in the presence of dislocations are discussed in appendix A5.1. The bifurcation set or geometrical caustic is shown dashed on these plots. It is derived by solving $\phi_1 = \phi_2 = 0$, from which x can easily be eliminated in this simple case to give

$$W_2^2 + 4W_1^3/27 = 0 \tag{5.3}$$

The method used to plot this equation as a dashed line is also described in appendix A5.1. "Inside" the caustic there are 3 SPPs or rays, "outside" there is only 1, and we see from fig.5.1 that the inside is considerably "brighter" than the outside. The caustic is obviously symmetrical in W_2 and so is $C(W_1,W_2)$ as seen by making the transformation $x \to -x$.

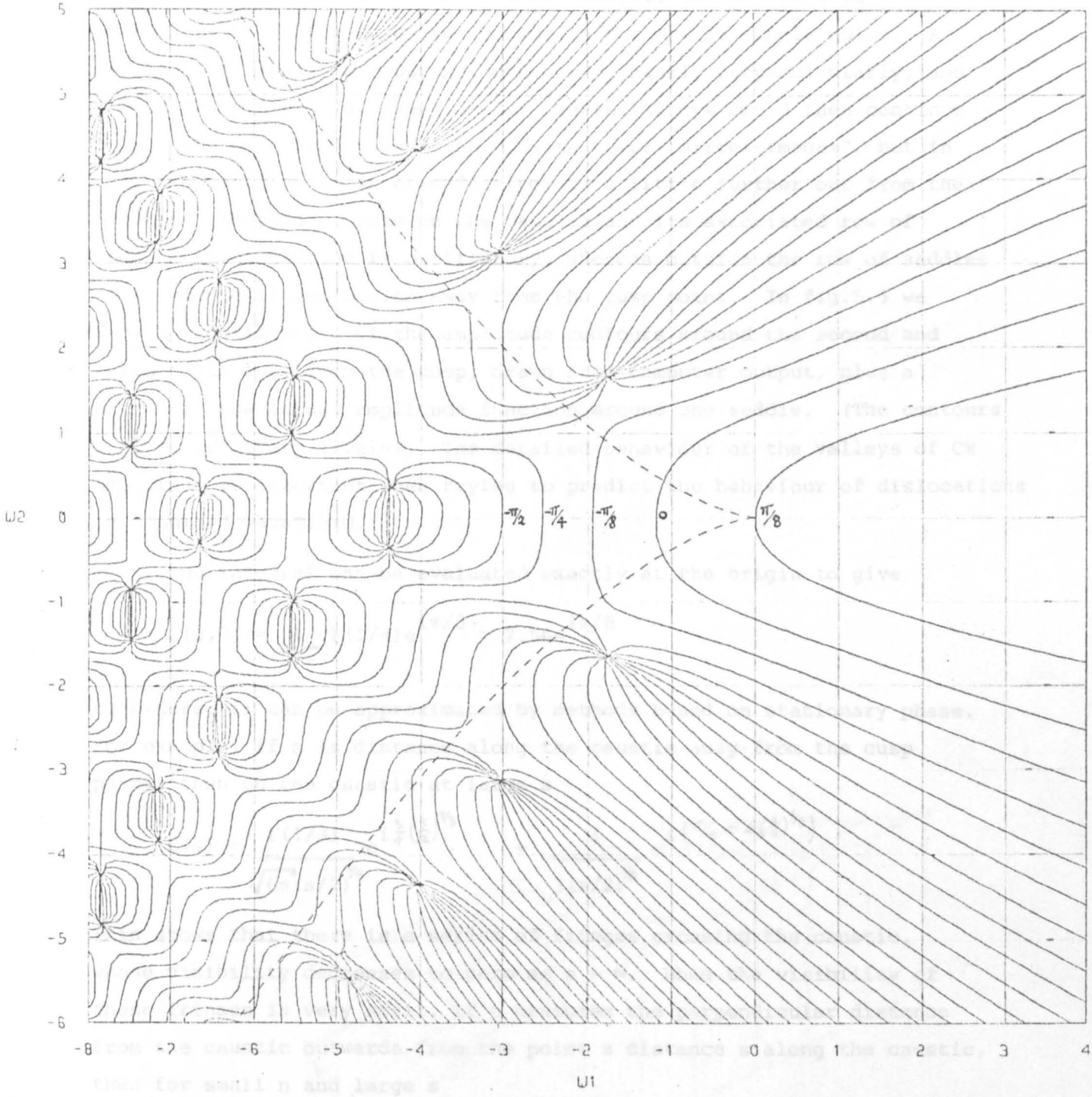
Our plots agree well with those of Pearcey(1946). His plots are interesting because they were produced mainly by integrating the partial differential equation satisfied by $C(W_1,W_2)$, using the Cambridge differential analyser. This was a mechanical analogue computer, and part of one may be seen in the Science Museum in London. This method of solution is ideally suited to such a computer, whose input and output devices are graphical, but is not the best method on a digital computer. Anyone who has seen a differential analyser, the prototype of which was built from "Meccano", cannot fail to be impressed by the accuracy of Pearcey's plots, compared with those produced by today's immensely powerful computers. Berry(1975) and Trinkaus & Drepper(1977) show plots of the intensity of the cusp diffraction pattern.

The most striking feature of the cusp diffraction pattern is the triangular array of pairs of nulls inside the caustic, and the row of single nulls just outside the caustic. We shall discuss these in more detail in the next section. The triangular array of bright spots inside



CUSP
CONTOURS OF AMPLITUDE

Figure 5.1



CUSP
CONTOURS OF PHASE IN Pi/8 STEPS

Figure 5.2

the caustic is what one would expect from two overlapping sets of Airy fringes. Outside the caustic the amplitude decreases away from the main maximum just inside the cusp point. Well away from the caustic this decrease is monotonic, but near to the caustic the decrease is oscillatory, as it must be to accomodate the row of single nulls. A "valley" of amplitude crosses the caustic, nearly perpendicularly, and passes through each of these nulls. These valleys could then continue around the main summit and link up with their "mirror-images", but in fact they stop on a row of saddle points a little further out from the caustic. Therefore, between the caustic and its associated row of saddles the amplitude is oscillatory, whereas outside the row of saddles it is monotonic decreasing away from the cusp point. In fig.5.3 we show a magnification of the amplitude contours around the second and third nulls away from the cusp, drawn from computer output, plus a sketch of the actual amplitude function around one saddle. (The contours stop at arbitrary levels). The detailed behaviour of the valleys of CW amplitude is important when trying to predict the behaviour of dislocations in a pulsed wavefield.

The integral can be evaluated exactly at the origin to give

$$C(0,0) = \frac{1}{2\sqrt{\pi}}\,\Gamma(1/4)e^{i\pi/8} \simeq 2.56e^{i\pi/8}$$

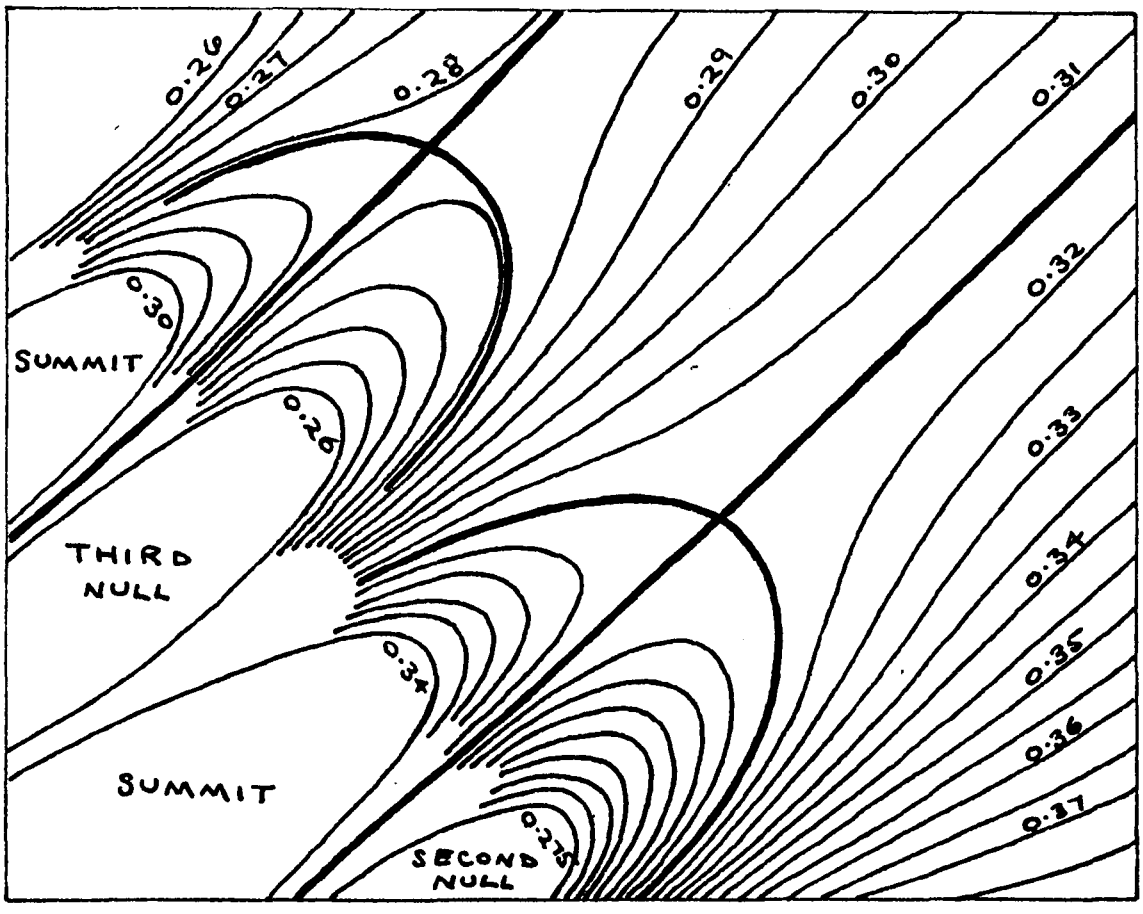
Elsewhere, it can be approximated by methods based on stationary phase. For example, if s is distance along the caustic away from the cusp point, then on the caustic at large s

$$C(s) \simeq \frac{\Gamma(1/3)}{\sqrt{6\pi}(s/2)^{1/9}}\,e^{i\frac{3}{4}\left(\frac{s}{2}\right)^{4/3}} + \frac{1}{3(s/2)^{1/3}}\,e^{i\left(\pi/4 - 6\left(\frac{s}{2}\right)^{4/3}\right)}$$

This shows that there is a series of fringes crossing the caustic, whose visibility decreases to zero as $s \to \infty$. When the visibility of these fringes is very small, if n measures the perpendicular distance from the caustic outwards from the point a distance s along the caustic, then for small n and large s

$$C(s,n) \sim \frac{\sqrt{2\pi}}{3^{1/3}}\,\frac{\mathrm{Ai}\left((s/2)^{2/9}n/3^{1/3}\right)}{(s/2)^{1/9}}\,e^{i\frac{3}{4}\left(\frac{s}{2}\right)^{4/3}}$$

This represents Airy fringes roughly parallel to the caustic, whose spacing and <u>intensity</u> are proportional to $(1/s)^{2/9}$.



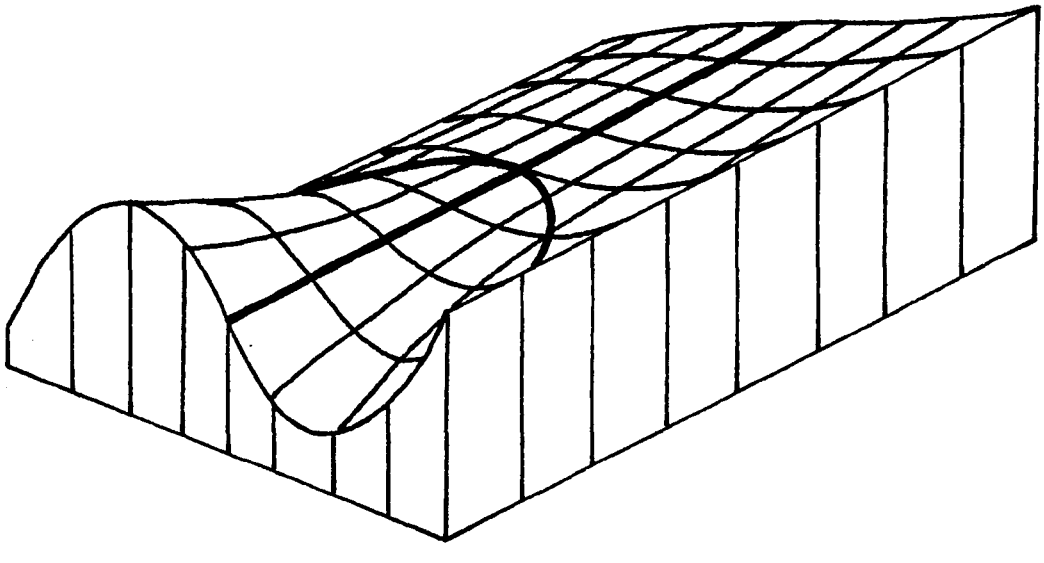


Figure 5.3 Saddle Points Outside the Caustic
(contours stop at arbitrary levels)

5.2 Nulls by Stationary Phase

We already know that the method of stationary phase predicts one sort of wavefield singularity, the caustic, but does it also predict the positions of the nulls with any accuracy? The one-dimensional stationary phase formula is

$$\int_{-\infty}^{\infty} dx\, e^{i\phi(x)} \simeq \sum_i \sqrt{\frac{2\pi}{|\phi_2(x_i)|}}\, e^{i\phi(x_i)}\, e^{i(\mathrm{sgn}\phi_2(x_i))\pi/4} \tag{5.4}$$

where $\phi_1(x_i) = 0$. So we could write down an approximation to the integral and then look for its zeros. But the x_i appearing in the above formula are the roots of a polynomial, and generally no simple formula for these exists, so the expression whose zeros we seek can be very complicated. We therefore introduce a further approximation, and proceed by "perturbing" a case for which we can find the SPPs easily.

The method is based on the fact that the SPPs $\underline{x}_i$ are the roots of polynomials whose coefficients are essentially the components of $\underline{W}$. Then away from any caustics, $\underline{x}_i$ is a smooth continuous function of $\underline{W}$, which can be approximated by its tangent, i.e. $\delta\underline{x}_i = O(\delta\underline{W})$. Suppose $\underline{x}_i = 0$ at $\underline{W} = 0$, i.e.

$$\nabla_{\underline{x}}\phi(\underline{0},\underline{0}) = 0.$$

Then for small $\underline{W}$ and $\underline{x}$

$$\phi(\underline{W},\underline{x}) = \phi_o + \underline{x}.\nabla_{\underline{x}}\phi_o + \underline{W}.\nabla_{\underline{W}}\phi_o + \tfrac{1}{2}(\underline{x}.\nabla_{\underline{x}} + \underline{W}.\nabla_{\underline{W}})^2\phi_o + \ldots \tag{5.5}$$

where $\phi_o \equiv \phi(\underline{0},\underline{0})$, so

$$\phi(\underline{W},\underline{x}) = \phi_o + \underline{W}.\nabla_{\underline{W}}\phi_o + O(\underline{W}^2) \tag{5.6}$$

because $\nabla_{\underline{x}}\phi_o = 0$ and $\underline{x} = O(\underline{W})$. ϕ_2 will also vary linearly in $\underline{W}$, but we neglect the linear variation of amplitude which this produces compared with the exponential variation produced by ϕ (unless the amplitude variation becomes important because of symmetry, for example, as occurs for the swallowtail).

Near a simple caustic (fold), however, catastrophe theory tells us that $\phi(x) = x^3/3 + Wx$ (up to diffeomorphism). Now $x_i = \pm\sqrt{-W}$, which is non-analytic at $W = 0$ (this is the catastrophe singularity!), and the orders of the terms of (5.5) are changed. Also $\phi_2 = \pm 2\sqrt{-W}$, and this variation is essential, producing the infinite amplitude on the caustic.

Along the W_1 axis of the cusp pattern the SPPs are given very simply by

$$\phi_1 = x^3 + W_1 x = x(x^2 + W_1) = 0$$

to be $x_i = 0, \pm\sqrt{-W_1}$. We shall initially consider only the region inside the cusp, away from the cusp point where $W_1 < 0$, where there are 3 real SPPs. Eqn.(5.6) gives

$$\phi(x(W_1,W_2),W_1,W_2) = \phi(x_i,W_1,0) + W_2 x_i + O(W_2^2)$$

$$= x_i^4/4 + W_1 x_i^2/2 + W_2 x_i + O(W_2^2)$$

and $\phi_2(x(W_1,W_2),W_1,W_2) = 3x_i^2 + W_1 + O(W_2)$.

Then $x_i = 0$ gives $\phi \simeq 0$, $\phi_2 \simeq -(-W_1)$

and $x_i = \pm\sqrt{-W_1}$ gives $\phi \simeq -(-W_1)^2/4 \pm W_2\sqrt{-W_1}$, $\phi_2 \simeq 2(-W_1)$.

Putting these values in (5.4) gives

$$C(W_1,W_2) \simeq \frac{e^{-i\pi/4}}{\sqrt{-W_1}} \left\{ 1 + e^{i(\pi/2 - W_1^2/4)\sqrt{2}} \cos W_2\sqrt{-W_1} \right\} \qquad (5.7)$$

for $W_1 < 0$, W_2 small. Zeros require either

(a) $\cos W_2\sqrt{-W_1} = -1/\sqrt{2} \Rightarrow W_2\sqrt{-W_1} = \pm\pi/4 + (2N+1)\pi$

and $\pi/2 - (-W_1)^2/4 = -2M\pi \Rightarrow W_1 = -\sqrt{(4M+1)2\pi}$

or (b) $\cos W_2\sqrt{-W_1} = +1/\sqrt{2} \Rightarrow W_2\sqrt{-W_1} = \pm\pi/4 + 2n\pi$

and $\pi/2 - (-W_1)^2/4 = \pi - 2m\pi \Rightarrow W_1 = -\sqrt{(4m-1)2\pi}$.

Obviously, (a) and (b) type solutions alternate to give a "brickwork" stacking of null pairs. The theory does not tell us how many null pairs there are for each W_1, but we know that we must not go too near the caustic. In the following table we calculate the coordinates of the first few nulls with $W_2 > 0$, and compare them with the values measured from figs.5.1 & 5.2. The stationary phase values in brackets are too near the caustic to actually occur. Essentially this analysis was first performed by Berry. It is included here for completeness, and to illustrate our general method, which will be used again later.

This formulation shows clearly why we get pairs of nulls inside the caustic. The two "side rays" combine to produce the cos term in (5.7). In the neighbourhood of each null pair these two rays can interfere in two symmetrical ways to produce a net amplitude of 1.

	Stationary Phase		Computed Graphs	
	W_1	W_2	W_1	W_2
a	-2.51	(1.49, 2.48)		
b	-4.34	0.38 (2.64, 3.39)	-4.40	0.35
a	-5.60	1.00, 1.66 (3.65, 4.31)	-5.55	1.00, 1.65
b	-6.63	0.30 2.13, 2.74 (4.57, 5.18)	-6.65 -6.45	0.30 2.15, 2.75

This resultant then interferes destructively with the third "central" ray at two nearby points. It is essential to have three interfering rays to produce these null pairs. We see that even this crude theory predicts the central nulls with very good accuracy, and even for quite large W_2 it is still accurate to about 3%.

There remains the row of single nulls just outside the caustic. Can these also be explained in terms of geometrical rays and stationary phase ideas? We can easily find the SPPs actually on the caustic because we know that one of the roots is double, so we shall proceed by "perturbing" this case. Consider the branch of the caustic defined by

$$W_{20} = +2\sqrt{(-W_{10}/3)^3}$$

where the subscript 0 indicates values on the caustic. Then the SPPs are

$$x_0 = +\sqrt{-W_{10}/3} \text{ (twice)} \quad \text{and} \quad x_0 = -2\sqrt{-W_{10}/3}$$

Now let us move off the caustic in the W_1 direction a small distance ξ to the point $W_1 = W_{10} + \xi$, $W_2 = W_{20}$. For perturbation about a caustic we cannot use the simple formula (5.6) for the double SPP, but must actually find the shift in the SPPs to lowest order. Suppose the SPP moves to $x = x_0 + \delta$, then we know that $\delta = O(\xi^{\frac{1}{2}})$. We find δ by expanding ϕ_1 about the caustic to lowest order:

$$\phi_1 = \phi_{10} + \delta\phi_{20} + \xi\partial\phi_{10}/\partial W_1 + \tfrac{1}{2}\delta^2\phi_{30} + O(\xi^{3/2}).$$

But $\phi_{10} = \phi_{20} = 0$, so requiring $\phi_1 = 0$ gives

$$\xi x_0 + \tfrac{1}{2}\delta^2 6x_0 + O(\xi^{3/2}) = 0$$

The SPP x_0 then drops out to give

$$\delta = \pm\sqrt{-\xi/3} + O(\xi).$$

Moving off the caustic has separated the two coincident SPPs. If we move off the caustic outwards, ξ is positive and δ is imaginary, so outside the caustic the double SPP becomes a pair of complex conjugate SPPs.

We can now expand ϕ and ϕ_2 about the caustic, remembering that $\delta = O(\xi^{\frac{1}{2}})$:

$$\phi = \phi_0 + \cancel{\delta\phi_{10}} + \xi\partial\phi_0/\partial W_1 + \cancel{\delta^2\phi_{20}}/2 + \delta\xi\partial\phi_{10}/\partial W_1 + \delta^3\phi_{30}/6 + O(\xi^2)$$

and $\phi_2 = \cancel{\phi_{20}} + \delta\phi_{30} + O(\xi)$. (Cancelled terms are actually zero.)

Putting $x_0 = \sqrt{-W_{10}/3}$ and $\delta = \pm i\sqrt{\xi/3}$ gives

$$\phi = -\frac{5}{4}\left(\frac{-W_{10}}{3}\right)^2 + W_{20}\left(\frac{-W_{10}}{3}\right)^{\frac{1}{2}} + \frac{\xi}{2}\left(\frac{-W_{10}}{3}\right) \pm i2\left(\frac{\xi}{3}\right)^{3/2}\left(\frac{-W_{10}}{3}\right)^{\frac{1}{2}}$$

and $\phi_2 = \pm i2(-W_{10}\xi)^{\frac{1}{2}}$

For the single SPP $x_0 = -2\sqrt{-W_{10}/3}$ we can use (5.6) to expand about the caustic, giving

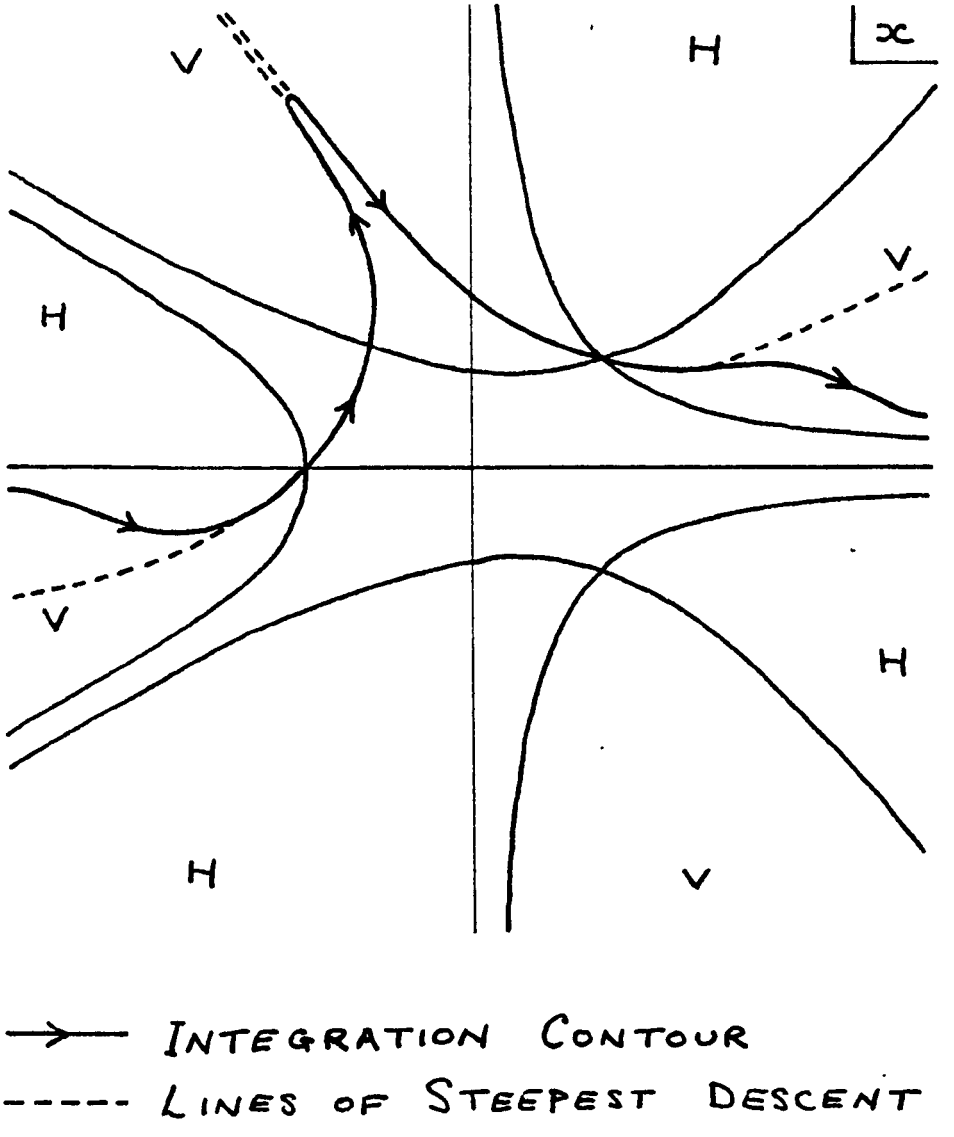
$$\phi = \phi_0 + \xi\partial\phi_0/\partial W_1 + O(\xi^2)$$

and $\phi_2 = \phi_{20} + O(\xi)$

i.e. $\phi = -2(-W_{10}/3)^2 - 2W_{20}(-W_{10}/3)^{\frac{1}{2}} - 2\xi W_{10}/3$

and $\phi_2 = -3W_{10}$

Because we now have complex SPPs we must use the method of steepest descents (Dennery & Krzywicki (1967), sec.31), which is a generalization of the method of stationary phase to complex stationary points. In fig.5.4 we show the "topography" of $|e^{i\phi(x)}| = e^{-Im\phi(x)}$ just outside the caustic. Note that $e^{-Im\phi(x^*)} = e^{+Im\phi(x)}$. H indicates a hill, where the integrand becomes infinite asymptotically, and V indicates a valley, where the integrand becomes zero asymptotically. These are demarkated by the contour lines passing through the saddle points where ϕ is stationary. Originally, the contour of integration passed along the real axis. We now distort it in the finite x plane so that it runs along the lines of steepest descent passing up to a saddle and then down the other side, but remains asymptotically along the real axis.



→ INTEGRATION CONTOUR

----- LINES OF STEEPEST DESCENT

Figure 5.4 Topography of the Cusp Integrand Outside the Caustic

The way this is done is forced upon us by the topography. The distorted contour is shown in fig.5.4, and we see that only one of the pair of complex saddles is used. (The only way to use the second complex saddle would involve traversing effectively the same contour twice in opposite directions, so that it would cancel out).

The integrand along the contour is now maximal at the two contributing saddles. Therefore we expand the integrand about these two saddles, as we have already done, to second order and change variable to a coordinate along the contour. We can then do the Gaussian integrals, to give an approximation to $C(W_1,W_2)$. The method can be made exact in principle by developing an asymptotic series, but we shall not do so here. Adding the contributions from the real saddle near $x = -2\sqrt{-W_{10}/3}$ and the upper complex saddle at $x = +\sqrt{-W_{10}/3} + i\sqrt{\xi/3}$ respectively, gives

$$C(W_{10}+\xi,W_{20}) \simeq e^{i\frac{2}{3}\left(-(-W_{10})^2 + \xi(-W_{10})\right)} \frac{e^{i\pi/4}}{\sqrt{-3W_{10}}} + \frac{e^{\left(i\{\frac{3}{4}(-W_{10}/3)^2 + \frac{\xi}{2}(-W_{10}/3)\} - 2\left(\frac{\xi}{3}\right)^{3/2}(-W_{10}/3)^{1/2}\right)}}{2^{1/2}(-W_{10}\xi)^{1/4}} \tag{5.7}$$

for small ξ, where $W_{20} = 2\sqrt{(-W_{10}/3)^3}$.

Because the contour of integration lies along the real axis, all real SPPs contribute to the diffraction integral, and so represent geometrical rays. We can generalize this notion to include "complex rays", which are those complex SPPs which contribute to the diffraction integral. Then we see that for the cusp there is only one complex ray outside the caustic, although there are two complex roots of $\phi_1(x) = 0$, and one real ray. This is why the nulls outside the cusp are only single. Also the amplitude of the complex ray decays exponentially away from the caustic, so that only close to the caustic is it large enough to cancel the real ray. This is why there is only one row of nulls outside the caustic.

Quantitatively, the conditions that $C = 0$ are, from (5.7), the amplitude condition

$$\frac{1}{\sqrt{-3W_{10}}} = \frac{e^{-\frac{3}{4}(-W_{10}\xi^3)^{1/2}}}{2^{1/2}(-W_{10}\xi)^{1/4}} \tag{5.8}$$

which has precisely one solution for ξ, and the phase condition

$$2\left(-(-W_{10})^2+(-W_{10}\xi)\right)/3 + \pi/4 = 3(-W_{10}/3)^2/4 + (-W_{10}\xi)/6 - (2n+1)\pi \tag{5.9}$$

Raising (5.8) to the 6th power gives

$$\frac{2^3(-W_{10})^{3/2}\xi^{3/2}}{3^3(-W_{10})^3} = e^{-\frac{4}{3}(-W_{10}\xi^3)^{1/2}}$$

Putting $X = \frac{4}{3}(-W_{10})^{1/2}\xi^{3/2}$ we can write this as

$$X = \ln(1/X) + \ln(9(-W_{10})^2/2) \qquad (5.10)$$

which gives a convergent iteration scheme if $X > 1$. Solving this for a range of values of W_{10} between -1 and -10 shows that ξ is very close to 1 and almost independent of W_{10}. Eqn.(5.9) is a quadratic in $(-W_{10})$:

$$\tfrac{3}{4}(-W_{10})^2 - \tfrac{1}{2}(-W_{10}) - (2n+5/4)\pi = 0$$

which we can easily solve with $\xi = 1$ for a range of values of n. We can then use (5.10) to find more accurate values of ξ. We could then iterate the process, but this is not necessary. A single iteration gives the following results, which are compared with values measured from figs. 5.1 & 5.2.

Stationary Phase		Computed Graphs	
W_{10}	ξ	W_{10}	ξ
-2.65	1.11	-2.60	0.9
-4.04	1.11	-4.05	1.0
-5.03	1.11	-5.05	1.0
-5.85	1.10	-5.90	1.0

Considering the imprecision with which measurements from the graphs can be made, and the crudeness of the approximations, the results are surprisingly good. So we see that methods based on the stationary phase approximation plus simple perturbative solutions can give the positions of the nulls of the cusp diffraction pattern with very good accuracy, both inside and outside the caustic. The real significance of this fact becomes apparent when the methods are applied to three-dimensional diffraction patterns, as we shall do in great detail for the elliptic umbilic. Then the null lines are complicated twisted space curves which are not at all easy to deduce from a set of plane sections. The fact that we have achieved such quantitatively good results in this simple case justifies applying the methods to more interesting problems.

5.3 An Ultrasonic Cusp Diffraction Pattern

An ultrasonic experiment has been set up by V.F.Humphrey to examine the cusp diffraction pattern produced by a pulsed source. With ultrasound the wavelength is long enough to produce a diffraction pattern on a large scale, which is easily amenable to measurements. A reflecting surface having approximately the form of a parabolic cylinder was used, and a pair of transducers, as close together as possible, were moved in a plane perpendicular to the axis of the reflector and equidistant from its ends. We shall model this experiment by an infinite parabolic cylinder reflector having equation $z = ax^2$, and we shall represent the pair of transducers by a coincident uniform point source and uniform point receiver. The received signal is given approximately by Berry's (1972) formula, which we rederive for sound waves in appendix A5.2. There we also indicate how the effects of the polar diagrams of the source and receiver might be taken into account, but we shall not do so in the following analysis. Then our initial formula for the wavefunction is

$$\Psi(z,t) \simeq \frac{1}{2\pi c} \int d^2\underline{R}\, \frac{F'(t-2r(R,z)/c)}{r(\underline{R},z)(z-f(\underline{R}))}$$

(see fig.5.5a) where F(t) is the transmitted signal.

Before considering diffraction, obviously the main contributions to the reflected signal will come from points where the reflector is perpendicular to the "line of sight", i.e. the rays are the normals to the surface. The rays will focus at the centres of curvature of the surface, so that the geometrical caustic is the surface of centres of the reflecting surface, or in two-dimensions the evolute of the parabola. This is a cusp pointing down, with its point at the principal focus at $z = 1/2a$, whose equation is

$$(z - 1/2a)^3 = 27x^2/16a \qquad (5.11)$$

(see fig.5.5b). This agrees with the fact, noted in the appendix, that this situation is equivalent to launching waves of the same shape as the reflecting surface. Then near the cusp point only the part of the reflector directly below the transducer is contributing significantly to the received wavefunction, so that $f(\underline{R}) \ll z$ and Berry's formula is valid. This may cease to be the case if we move too far away from the cusp point, as discussed in appendix A5.2.

Let us transmit a monochromatic signal $F(t) = e^{-i\omega t}$, then putting $\underline{R} \equiv (x,y)$ and $k = \omega/c$ we have

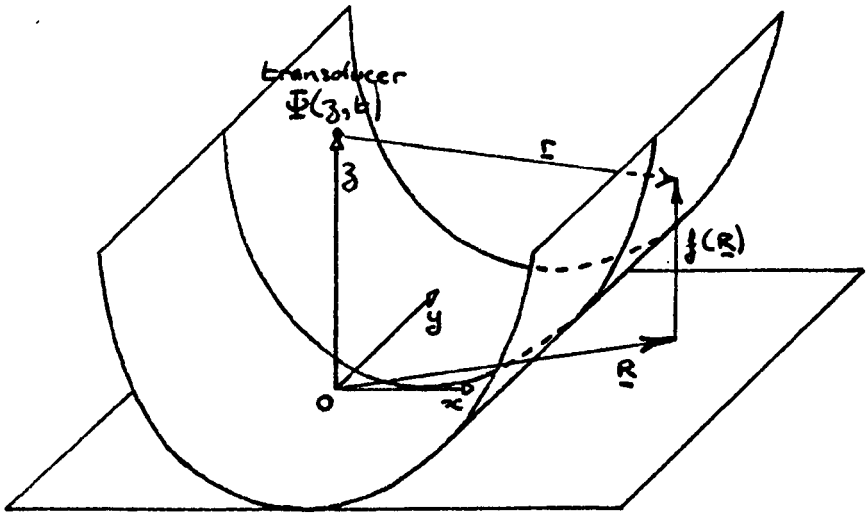


Figure 5.5a Infinite Parabolic Cylinder Reflector

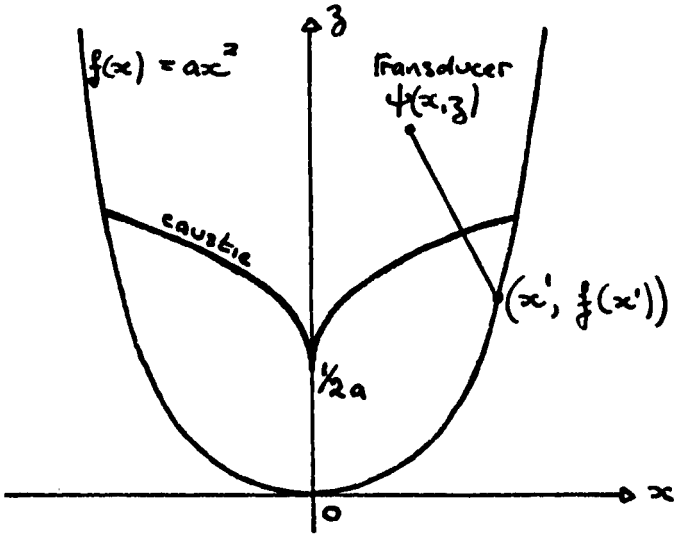


Figure 5.5b Two-Dimensional Section showing Caustic and New Origin

$$\Psi(z,t) \simeq -\frac{i\omega e^{-i\omega t}}{2\pi c} \int_{-\infty}^{\infty} \frac{1}{z-f(x)} \left(\int_{-\infty}^{\infty} dy \, \frac{e^{i2kr(x,y,z)}}{r(x,y,z)} \right) dx$$

We have assumed $kr >> 1$ already in the derivation of Berry's formula, therefore we should be able to evaluate the y integral reliably by stationary phase. We have

$$r(x,y,z) = \sqrt{y^2 + \{x^2 + (z-f(x))^2\}}$$

and the SPP occurs at $y = 0$, which is an obvious condition for normals to the reflector. Then working with an origin fixed relative to the reflector, rather than relative to the transducer, the signal received at some point (x,z) at time $t = 0$ is

$$\psi(x,z) \simeq \frac{1}{2}\sqrt{\frac{k}{\pi}}\, e^{-i\pi/4} \int_{-\infty}^{\infty} dx' \, \frac{e^{i2k\sqrt{(x'-x)^2 + (z-f(x'))^2}}}{(z-f(x'))\left((x'-x)^2 + (z-f(x'))^2\right)^{1/4}} \qquad (5.12)$$

(see fig.5.5b). We shall refer to (5.12) as the "exact" diffraction integral. Its stationary phase points occur at

$$(x'-x) - (z-f(x'))f'(x') = 0 \qquad (5.13)$$

which is exactly the condition that (x,z) lies on the normal at x'. Our surface has the equation $f(x) = ax^2$, so (5.13) becomes

$$2a^2x'^3 + (1-2az)x' - x = 0 \qquad (5.14)$$

which has 3 (possibly complex) solutions. We wish to map (5.12) onto a canonical diffraction integral, and to preserve the stationary phase points we must obviously use a quartic ϕ, i.e. we must map (5.12) onto the canonical cusp integral (which we know anyway in this simple case).

Note that if our surface had the form $f(x) = ax^n$ near the origin, then (5.13) would have $(2n-1)$ solutions and we would have to use a ϕ of order $2n$. Therefore the caustics produced in this way by an analytic cylindrical reflecting surface can only be cuspoid catastrophes of even codimension, i.e. cusp, butterfly, etc. For example, to produce a pure fold caustic would require a surface of the form $f(x) = ax^{3/2}$ which has two sheets meeting at a cuspidal edge along $x = 0$, where the surface is non-analytic.

As discussed in section 4.2, the condition for two SPPs to coalesce is the vanishing of the Hessian of the phase, which in this one-dimensional case is just the second derivative with respect to x'. Applied to (5.14) this gives

$$6a^2x'^2 + (1-2az) = 0.$$

Solving this with (5.14) gives (5.11), the exact equation of the geometrical caustic. This is also exactly the form of the canonical cusp caustic, so one is led to wonder to what extent the diffraction pattern is just a scaled version of the canonical pattern, i.e. is the mapping from (z,x) in (5.12) to (W_1,W_2) in (5.1) linear?

We shall make a transitional approximation to (5.12) in the neighbourhood of the cusp point, where the three SPPs coalesce at $x' = 0$. We therefore expand the phase in (5.12) about $x' = 0$ up to fourth order and shift the origin to remove the cubic term, giving

$$2k\{z - 2axx' + a(1-2az)x'^2 + a^3x'^4\}$$

to lowest order in the coefficients. We approximate the denominator by its value of $(1/2a)^{3/2}$ at the cusp, and define a new variable of integration t by

$$x' = -t(1/8ka^3)^{1/4} \qquad \Rightarrow \qquad 2ka^3x'^4 = t^4/4$$

Then (5.12) becomes

$$\psi(x,z,t) \simeq (2ka^3)^{1/4} e^{-i(\omega t-2kz+\pi/4)}\, C(W_1,W_2) \tag{5.15}$$

where $W_1 = -(8ka)^{1/2}\,(z-1/2a)$

$W_2 = (32ak^3)^{1/4}\,x$

which should be valid near $x = 0$, $z = 1/2a$. We can easily show that the caustic (5.3) of $C(W_1,W_2)$ maps exactly onto the required caustic (5.11). (5.15) is our canonical approximation to the "exact" wavefunction (5.12). Note the exponential factor which gives the main phase variation of ψ, since it varies much more rapidly than the phase of $C(W_1,W_2)$ (except near the nulls). In fact, the equiphase lines of $C(W_1,W_2)$ curve the wrong way for wavefronts; $C(W_1,W_2)$ only represents the phase shifts due to the diffraction.

This is actually the experiment that our previous analysis of the piston radiator was designed to model. Therefore, we use the same figures as before giving $k \simeq 19\ cm^{-1}$. The value of a is most accurately determined from fitting the shape of the caustic, and not from the height of the principal focus above the reflector. This is because any reflecting surface which is one of the family of involutes of the cusped caustic will produce the same caustic and presumably similar diffraction patterns. Only one of these surfaces is the parabolic cylinder which we took as our model, and the experimental surface may well have actually been a nearby member of the family of involutes.

The value found by Humphrey in this way was $a \simeq 0.1\ cm^{-1}$ giving $(2ka^3)^{1/4} \simeq 0.44$, $(8ka)^{1/2} \simeq 3.90$ and $(32ak^3)^{1/4} \simeq 12.17$. Then the linear approximation to the mapping near the cusp point is

$$W_1 = -3.90\ (z-5.0)$$

$$W_2 = 12.17\ x$$

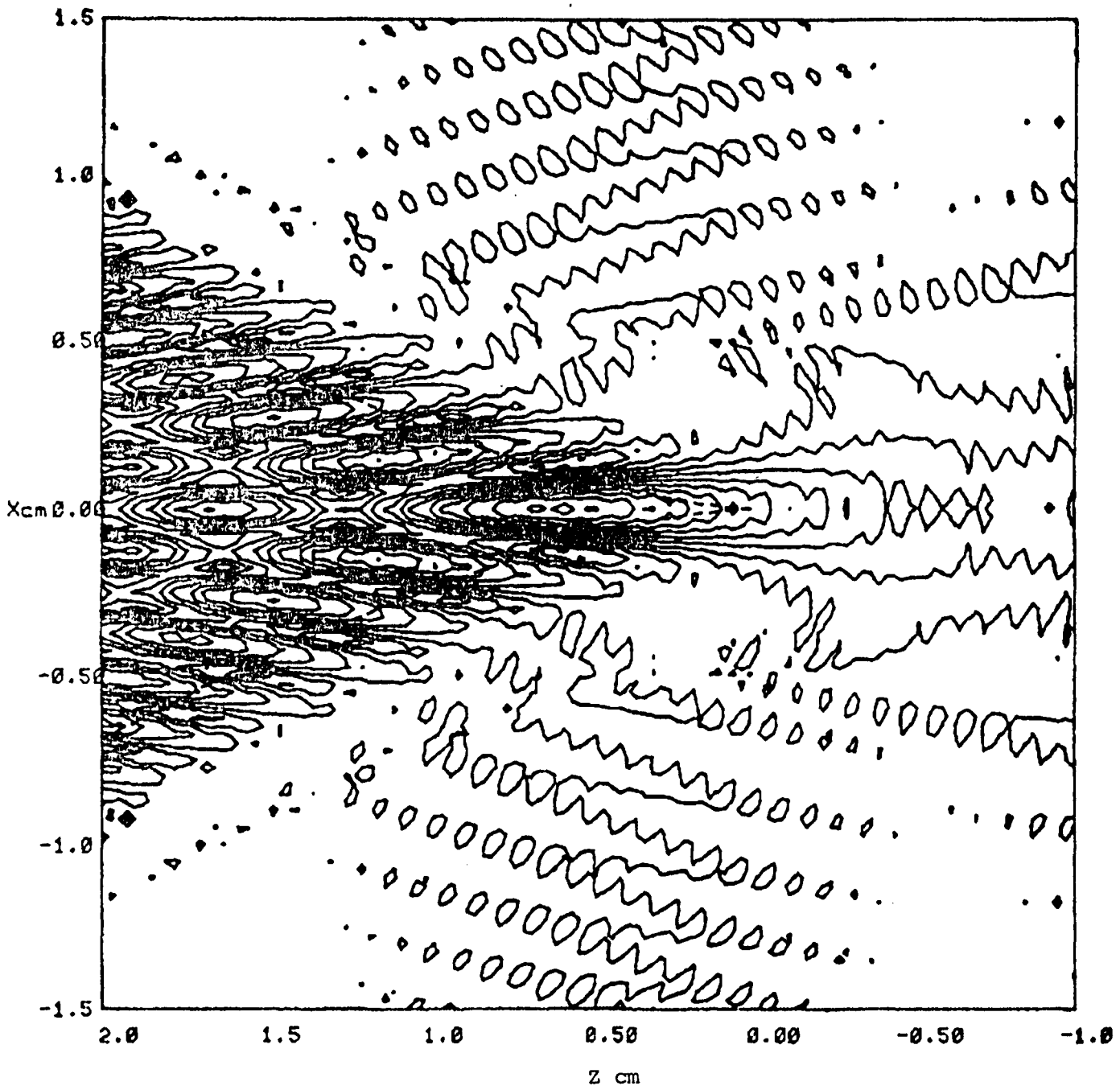
and the canonical pattern is squashed by a factor of about 3 in the x direction.

Note that the simple Berry diffraction formula is adequate for our transitional approximation, which is only valid near the cusp point anyway. However, the "exact" integral (5.12) is actually divergent, because our parabolic reflector violates the assumptions in the derivation of (5.12). The transitional approximation removes this divergence and it may be possible to compute (5.12) with care near the cusp point by removing the divergence "by hand"; however, we should beware that the results may not be reliable.

For comparison, we attempt to compute the cusp wavefunction from the "exact" formula (5.12) and from the transitional approximation (5.15), and display the results in figs.5.6 & 5.7 respectively. Additional computational details are given in appendix A5.1. The values at the cusp point agree very closely. The phase plots shown in figs.5.6b & 5.7b are remarkably similar, except that the overall scale of the canonical diffraction pattern is smaller than the exact pattern. Within a circle centred on the cusp point and passing through the closest nulls of the canonical pattern the phase plots match very closely. Outside this region the scale changes indicating that the linear mapping from (x,z) to $\underline{W}$ has broken down.

The amplitude plots shown if figs. 5.6a & 5.7a are less similar. The main difference is a "ripple" on the "exact" amplitude whose wavelength is exactly half the true wavelength. The fact that there is no sign of this ripple on the phase plot suggests that it is real, but it could be an artefact produced by trying to evaluate a divergent integral. Apart from this the "exact" amplitude is very similar to the approximate amplitude with the same non-linear scaling as the phase.

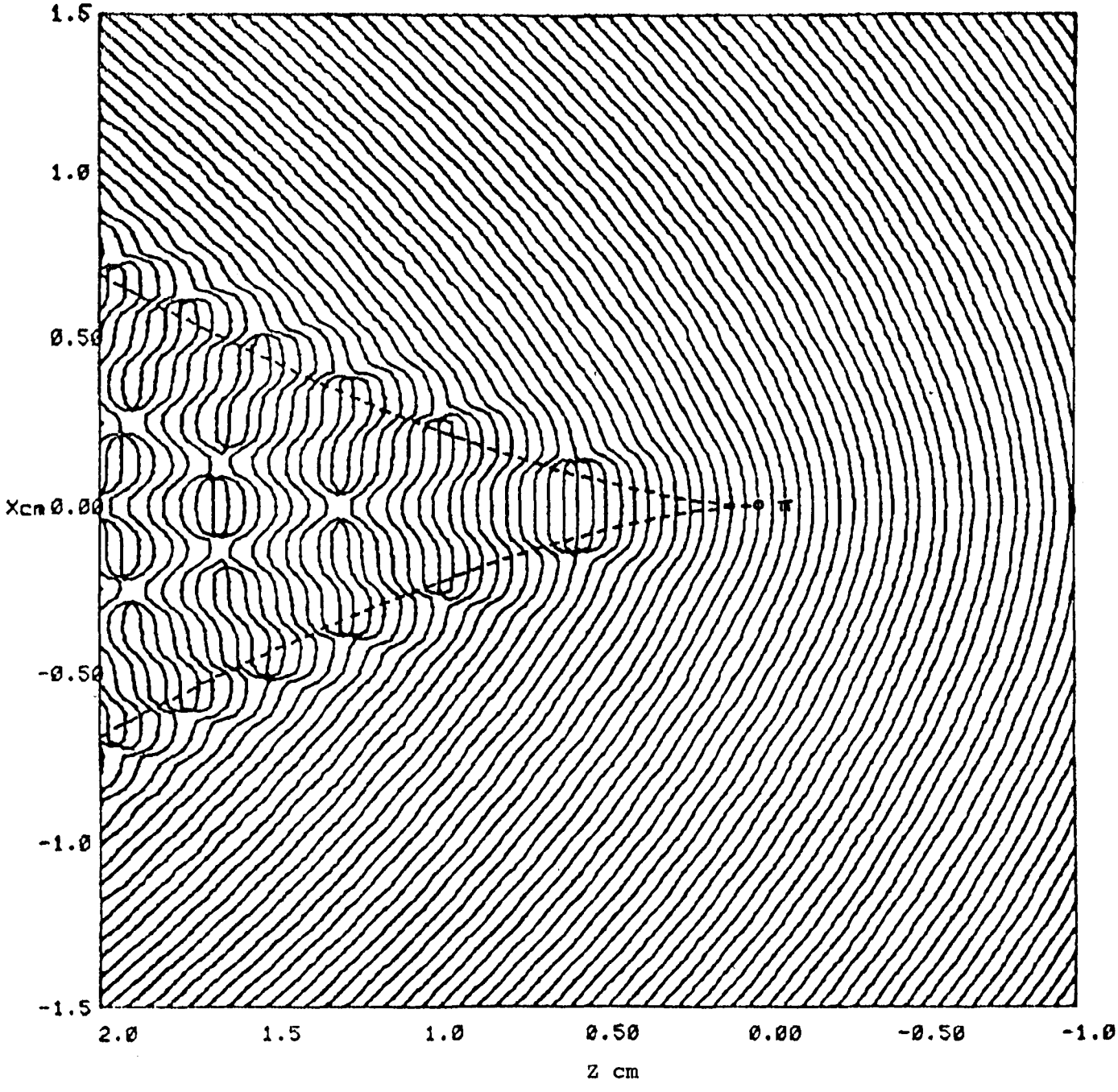
Our tentative conclusion, then, is that the simple canonical cusp diffraction pattern obtained by a transitional approximation is qualitatively correct, although the scaling is wrong away from the cusp point, and the apparent ripple on the exact wavefunction needs further investigation.



"EXACT" CUSP WAVEFUNCTION

CONTOURS OF AMPLITUDE

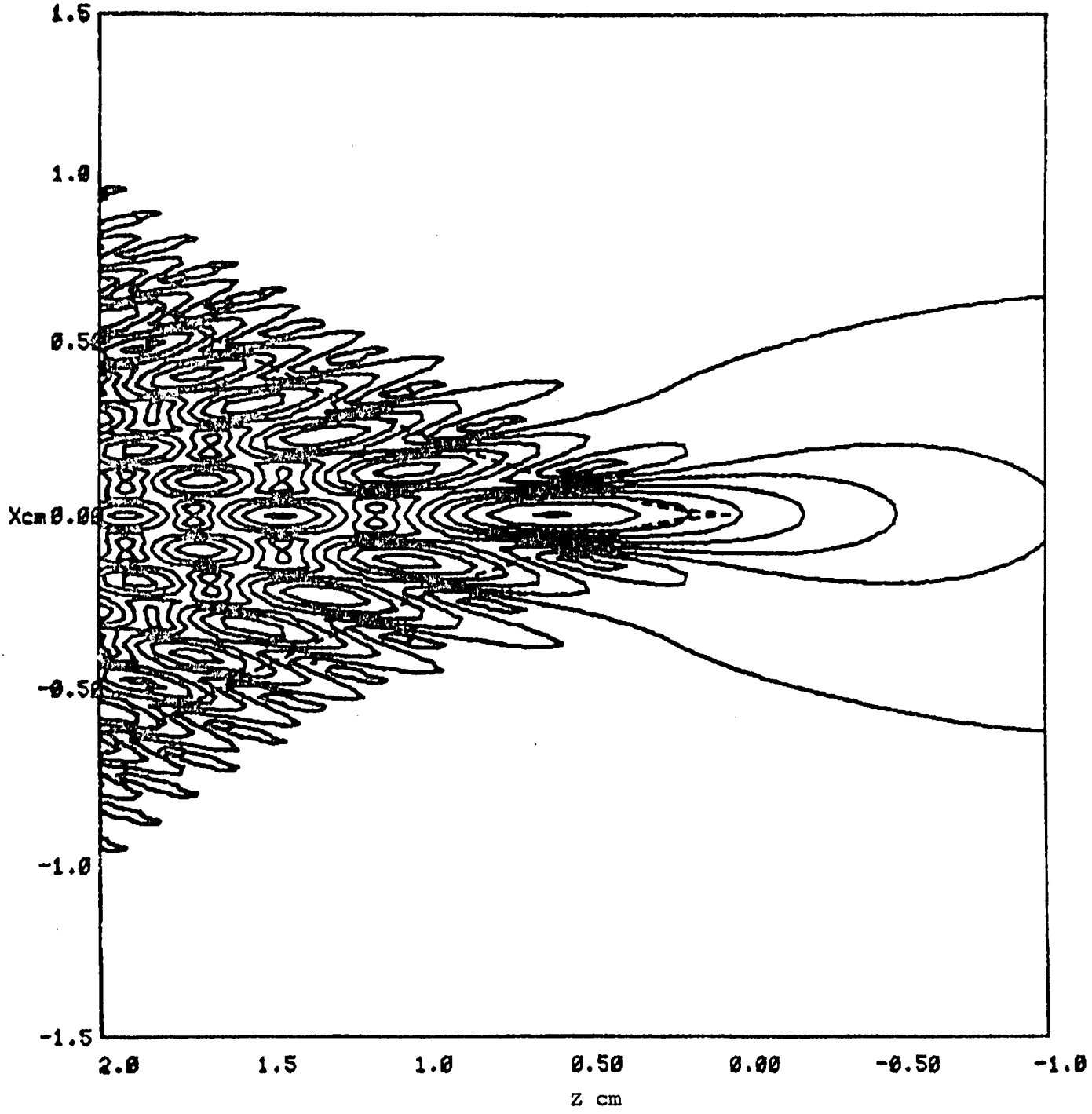
Figure 5.6a



"EXACT" CUSP WAVEFUNCTION

CONTOURS OF PHASE in steps of Pi/2

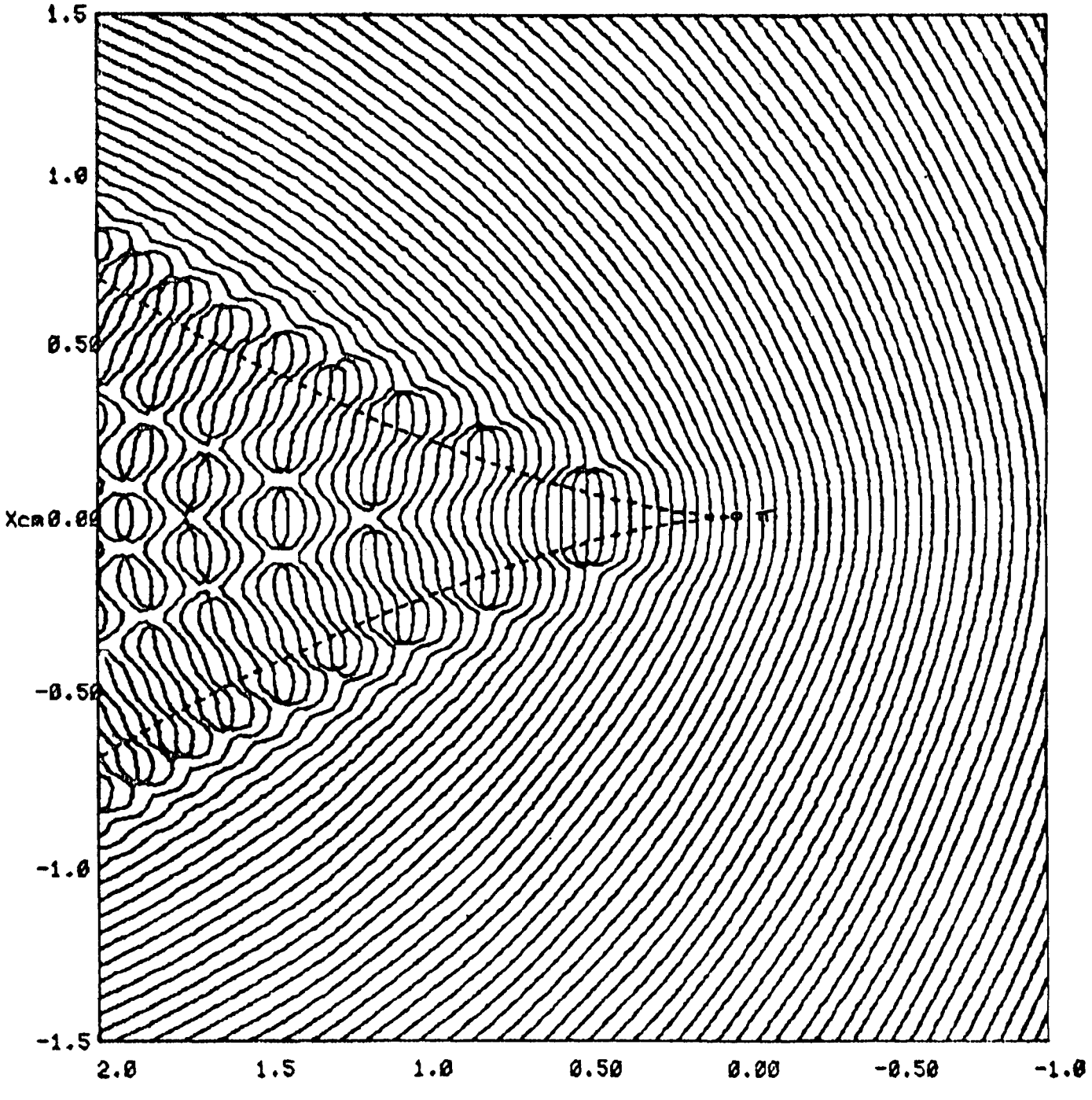
Figure 5.6b



CANONICAL APPROXIMATION TO CUSP WAVEFUNCTION

CONTOURS OF AMPLITUDE

Figure 5.7a



CANONICAL APPROXIMATION TO CUSP WAVEFUNCTION

CONTOURS OF PHASE in steps of Pi/2

Figure 5.7b

5.4 Dislocation Trajectories near the Cusp

So far we have only examined the CW cusp diffraction pattern, but as mentioned in the description of the experiment, the transducer was really driven by quasimonochromatic pulses. Therefore, we expect to find wavefront dislocations moving through a diffraction pattern which is otherwise similar to the CW pattern. In order to form an intuitive picture of the interaction of the pulsed rays, it is necessary to know the effect on a pulse of its ray touching a caustic and focussing, and this is discussed in appendix A5.3. We expect the dislocation trajectories to pass through, or very close to, the CW nulls and to run along the valleys and over the saddles of the CW amplitude. But looking at fig. 5.1 we see that there are many such paths which the trajectories could follow.

We found in chapter 3 that the trajectories are given to a good approximation by the frequency minima of the CW amplitude. We shall work with the canonical approximation (5.15). If we define

$$\mathcal{M}(W_1,W_2) = |C(W_1,W_2)|$$

then $M(x,z;\omega) = |\psi(x,z;\omega)| \propto \omega^{1/4}\mathcal{M}(W_1,W_2)$

where $W_1 = -(8a/c)^{1/2}(z-1/2a)\,\omega^{1/2}$

$$W_2 = (32a/c^3)^{1/4} \times \omega^{3/4}$$

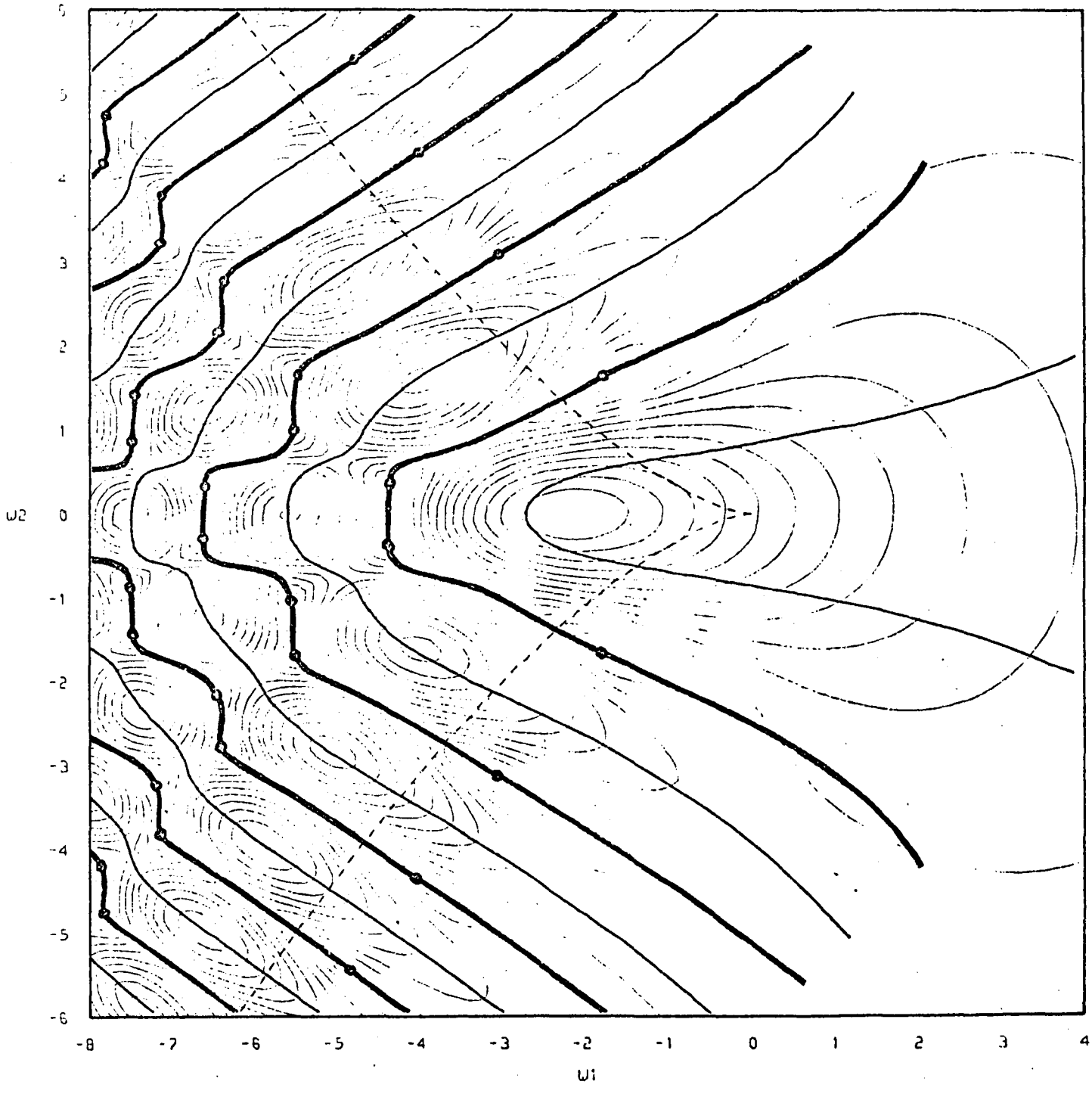
Then $$\frac{\partial M}{\partial \omega} = \omega^{1/4}\frac{\partial \mathcal{M}}{\partial \omega} + \frac{1}{4}\omega^{-3/4}\mathcal{M} = \omega^{1/4}\left\{\frac{W_1}{2\omega}\frac{\partial \mathcal{M}}{\partial W_1} + \frac{3W_2}{4\omega}\frac{\partial \mathcal{M}}{\partial W_2}\right\} + \frac{\omega^{-3/4}}{4}\mathcal{M}$$

So $\partial M/\partial\omega = 0$ implies

$$\mathcal{M} + 2W_1\frac{\partial \mathcal{M}}{\partial W_1} + 3W_2\frac{\partial \mathcal{M}}{\partial W_2} = 0 \qquad (5.16)$$

or $$\mathcal{M} + 2(z-1/2a)\frac{\partial \mathcal{M}}{\partial z} + 3x\frac{\partial \mathcal{M}}{\partial x} = 0$$

In this simple case we know explicitly the frequency dependence of the wavefunction, so that we can relate the frequency derivative to space derivatives, either in W_1W_2 space or in xz space. We shall work in W_1W_2 space for convenience. Then by numerically differentiating the values of $\mathcal{M}$ which we have previously computed, we can easily plot the solutions of (5.16). However, where $\mathcal{M}(W_1,W_2)$ is very flat the numerical differentiation is unstable. In fig.5.8 we plot the trajectories predicted by (5.16) (omitting the small unstable region) superimposed



CUSP

DISLOCATION TRAJECTORIES SUPERIMPOSED ON CW AMPLITUDE

Dark lines represent frequency minima, which pass through CW nulls shown as dots. Light lines represent frequency maxima.

Figure 5.8

on contours of $\eta(W_1,W_2)$. The curves fall into two classes. Those shown dark link up all the CW nulls by running up the valleys and over the saddles. Those shown light link up all the CW maxima and run down the ridges and over the saddles not used by the other curves. The curves all cross the caustic and run off towards the reflector.

Based on the discussion in chapter 3, we suggest that only the dark curves linking the CW nulls, are approximate dislocation trajectories and that the other set of curves is spurious. Note that near the null pairs the real trajectories have fairly square steps. We have already set up a local model of such symmetric pairs of stepped trajectories in section 1.6, which are illustrated in fig.1.20, and result from cubic terms in the local expansion of the wavefunction. These trajectories predicted near the cusp point agree very well with experiment.

We know that near to the caustic far from the cusp point, the diffraction pattern is just the fold or Airy pattern. This pattern is essentially one-dimensional, so the dislocation trajectories must run parallel with the Airy fringes and the caustic. This is the same as occurs in the far field of the piston radiator. We notice that as we move away from the cusp point, the angle at which the trajectories intersect the caustic increases from about 45° to about 90°. I suggest that this angle continues to increase until asymptotically the trajectories run parallel with the caustic. However, in the real experiment the time delay between the pair of pulsed rays which focus on the caustic and the third pulsed ray from the other side of the reflector becomes so great that they do not interact. Therefore the Airy limit in the experiment is approached much more rapidly than for continuous waves, and one would expect a more complicated transition between "cusp trajectories" and "fold trajectories".

No attempt has been made to predict the dislocation times, because the theory is not sufficiently accurate, at least not for a pulse as short as that used. Further research on this could be useful because the nulls of this continuous wavefunction are more general than the axial null of the piston radiator. The methods for handling CW nulls discussed in section 3.8 should be easier to put into practice for such general nulls, and it would be interesting to see if it can be done satisfactorily. It would also be interesting to try using a more exact diffraction integral to see if the ripple, if it really exists, has any effect on the trajectories.

APPENDICES

A5.1 Computational Details

The basic integration method is discussed in section 4.5. The first step is to examine the behaviour of

$$\phi_1(x,W_1,W_2) = x^3 + W_1x + W_2$$

which is sketched in fig.A5.1.1. We only need consider $W_2 \geq 0$. For some fixed W_2, for $W_1 \geq 0$ there is only one SPP at $x \leq 0$. As W_1 is decreased this SPP moves left. At some value of $W_1 \leq 0$ we cross the bifurcation set and a pair of SPPs appear at $x \geq 0$. As W_1 is further decreased these separate such that the rightmost one moves right. Therefore the outermost SPPs separate as W_1 is decreased. The stepping algorighm is: start at $W_2 = 0$, with A = +STEP, B = -STEP and W_1 = max. Increase A in increments of STEP until $\phi_1(A) \geq C1$. Decrease B in decrements of STEP until $\phi_1(B) \leq -C1$. Step A and B out for convergence of the asymptotic series. Set BS = B to use for the next W_1 scan. Decrease W_1. Step out A and B if necessary until $\phi_1(A) \geq C1$ and $\phi_1(B) \leq -C1$, and then for convergence. Decrease W_1 etc. until W_1 = min. Increase W_2 and set A = +STEP, B = BS and repeat until W_2 = max.

This algorithm will work for $W_2 < C1$, otherwise A will not be forced to jump outside the new pair of SPPs which appear on crossing the bifurcation set. This can be handled by setting $A = \sqrt{-W_1/3}$ if $\phi_1(\sqrt{-W_1/3}) < C1$, and then stepping out from there as necessary. Also if $W_2 > C1$ the value A = +STEP may be larger than necessary at the beginning of the W_1 scans, and could cause spurious convergence. We could solve the equation $\phi_1(AS) = C1$ to low accuracy by Newton's method and start the scan with A = AS, and then for the next scan use AS to start the Newton's iteration for a new AS. This would work for all W_2, and would probably be more efficient even for $W_2 < C1$.

For a series error of less than 0.0005 the asymptotic forms in section 4.5 give the cutoffs as ±2.63, and the magnitudes of the first 3 terms of the series as

0.05, 0.003 & 0.0005

It was found satisfactory in practice to use C1 = 10, and require the second term to be less than 0.01, and the third less than 0.0005. This should give a total series error of less than **0.001.**

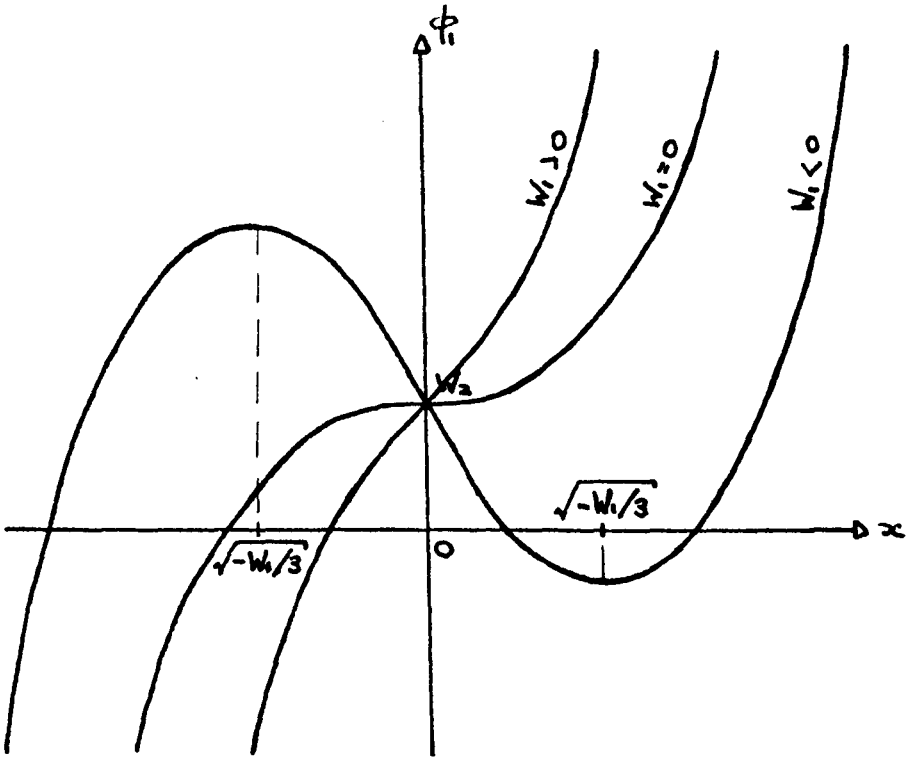


Figure A5.1.1 ϕ_1 for the Cusp Integral

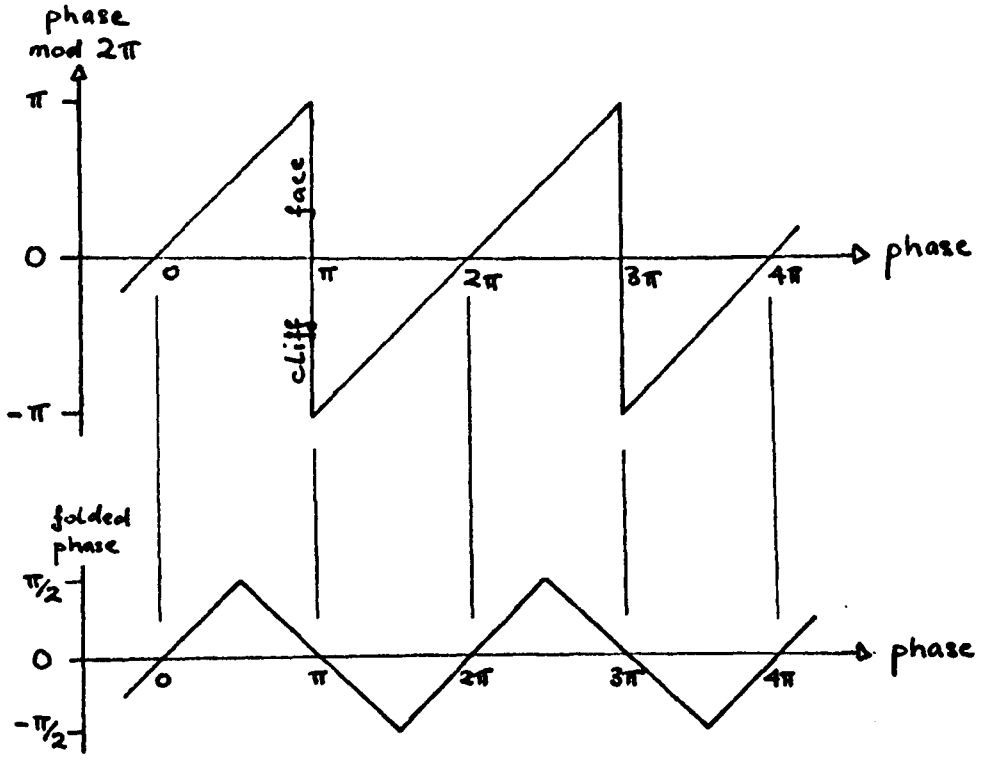


Figure A5.1.2 Folding the Phase

The "exact" cusp wavefunction (5.12) is computed by a generalization of the basic method. Changing the variable of integration to -x' improves the analogy with the canonical integral and gives

$$\psi = \frac{1}{2}\sqrt{\frac{k}{\pi}}\, e^{-i\pi/4} \int_{-\infty}^{\infty} dx' \; \frac{e^{i2k\sqrt{(x'+x)^2 + (z-ax'^2)^2}}}{(z-ax'^2)\{(x'+x)^2 + (z-ax'^2)^2\}^{1/4}} \qquad \text{(A5.1.1)}$$

This integrand is singular at $ax'^2 = z$. But our original derivation (appendix A5.2) assumes $z < ax'^2$ always, i.e. the reflecting surface is always below the transducer. In practice, because of the polar diagram of the transducers, the surface above the transducer plays no part, and we would be justified in only integrating over some range smaller than $(-\sqrt{z/a}, +\sqrt{z/a})$. In fact we only integrate numerically over a much smaller range, leaving the rest of the integral to be handled by the asymptotic series. But we must check that the cutoffs remain within the above allowed range.

Let us define

$$\$ = (x'+x)^2 + (z-ax'^2)^2$$

$$\$_1 = 2(x'+x) - 4ax'(z-ax'^2) = 2\{2a^2x'^3 + (1-2az)x' + x\}$$

$$\$_2 = 2 - 4az + 12a^2x'^2$$

$$Z = z-ax'^2$$

$$Z_1 = -2ax'$$

Then integrating by parts as in section 4.5 gives

$$\int_A^B dx' \; \frac{e^{i2k\$^{1/2}}}{Z\$^{1/4}} = \left[G(x) + O(1/k^3) \right]_A^B$$

where $G(x) = \frac{\$^{1/4} e^{i2k\$^{1/2}}}{Z\$_1} \left\{ \frac{-i}{k} + \frac{1}{k^2}\left(\frac{1}{4\$^{1/2}} - \frac{\$^{1/2}(Z\$_2 + Z_1\$_1)}{Z\$_1^2} \right) \right\}$

Now if we only use (A5.1.1) where the assumptions in its derivation are valid, then physically G(x) must be negligibly small at the physically meaningful upper and lower limits of the integral. Outside these limits, of course, it will diverge when Z = 0. We shall take G(x) = 0 at the upper and lower limits, and assume that this is justified within the xz region in which we are interested. This assumption needs further checking: a test of its validity should have been included in the program but was not, and consequently the results presented are only preliminary. If the assumption turns out to be invalid we must use formula (A5.2.2)

instead of (A5.2.3) as our starting point for the "exact" integral. The convergence criteria used were $|\$_1| > 1.0$, and the first and second terms of $G(x) < 0.02$ and 0.0005 in magnitude. The remaining details of the algorithm are essentially as for the canonical integral. Note that $G(x)$ is a simple asymptotic power series in k, and here $k = 19$ so it should converge rapidly.

We plot the caustic by a series of dashes of approximately equal lengths, whose ends lie exactly on the caustic curve. For $W_2 \geq 0$ the equation of the caustic is

$$W_2 = \frac{2}{3}\left(\frac{-W_1{}^3}{3}\right)^{\frac{1}{2}} \quad \Rightarrow \quad \frac{dW_2}{dW_1} = -\left(\frac{-W_1}{3}\right)^{\frac{1}{2}}$$

Then the arc length s is given by

$$ds/dW_1 = \pm\sqrt{1 - W_1/3}$$

and the W_1 coordinates of successive points along the curve a distance δs apart are given approximately by

$$W_{1,n+1} = W_{1,n} - \delta s/\sqrt{1 - W_{1,n}/3}$$

where $W_{1,0} = 0$. Then $W_{2,n+1} = \frac{2}{3}\left(\frac{-W_{1,n+1}{}^3}{3}\right)^{\frac{1}{2}}$ exactly, and we plot (W_1,W_2) and $(W_1,-W_2)$ to give the whole caustic. This method is easily generalized to the higher catastrophes where the caustic can easily be specified parametrically, but not as explicit equations.

There remains the problem of plotting equiphase lines in the neighbourhood of a dislocation. Phase is only defined modulo 2π, hence it is a discontinuous function. As a surface it has vertical cliffs, and if we try to plot contours of this, all possible contours will be drawn in this cliff face. If a wavefunction contains a dislocation in some region there is no way of removing all phase discontinuities in that region. But we know that phases differing by π become associated at a dislocation, therefore we can preprocess the phase data so that phases differing by π become the same, and produce a continuous function in the range $[-\pi/2,+\pi/2]$. To plot the contour phase = p, we first shift the phase to lie in the range $[p-\pi,p+\pi]$ so that the phase is smooth for an equal distance either side of p, to facilitate linear interpolation. We then "fold" the data about $\pm\pi/2$ using the algorithm

```
data:= if data > π/2 then π-data else
       if data < -π/2 then -π-data else data
```

The effect is shown symbolically in fig.A5.1.2.

Now by plotting the zero contour of this preprocessed data, we automatically plot both the equiphase lines at p and at p ± π. However, if the data calculated is actually $\psi = R + iI$, the whole complex wavefunction, then the phase is calculated from arctan(I/R) and may not be very accurate around R = 0. But we can easily plot the equiphase lines at $0, \pi$ and $\pm\pi/2$ by plotting the contours I = 0 and R = 0 respectively. To plot any other set of 4 equiphase lines we just rotate ψ in the complex plane and then plot I = 0 and R = 0. For example, to plot phase = $+\pi/4$, $-3\pi/4$ and $-\pi/4$, $+3\pi/4$ we plot R - I = 0 and R + I = 0 respectively.

Both methods have been used and both give good results. The method to use depends on the form of the data, but the second should be more efficient and probably gives slightly better results, although this will depend on the form of the wavefunction (and the sophistication of the contouring routine).

A5.2 Diffraction Integral for Point Sonic Source-Receiver

We wish to derive an expression for the signal received by a point receiver coincident with a point sonic radiator after reflection by a perfect rigid reflecting surface. This problem has been solved by Berry (1972) for electromagnetic waves, where the boundary condition is different. We follow Berry's method closely, but use the correct boundary condition for sound waves at a rigid interface, and end up with minus Berry's formula. We also consider the effect of a polar diagram for the source and receiver.

The reflecting surface Σ has height $f(\underline{R})$ above the reference plane in which $\underline{R}$ is a two-dimensional vector from the origin, as in fig.A5.2.1. The source-receiver S is a height h above the origin. $\underline{r}(\underline{R})$ is the vector from S to the point on Σ above $\underline{R}$, and $\hat{\underline{n}}$ is the inward normal at that point. Now the incident wavefunction $\psi_s(t)$ at the receiver is given by the Kirchhoff integral representation (e.g. see Jackson (1962), p188; note different conventions) as an integral of the incoming wavefunction over the closed surface Σ'. We can close Σ by a "hemisphere at infinity" on which there is no reflected wave, so that it does not contribute, giving

$$\psi_s(t) = \frac{-1}{4\pi}\int_{\Sigma} d\Sigma\ \hat{\underline{n}}.\left[\frac{\nabla\psi_{ref}(\underline{r},t)}{r} + \frac{\hat{\underline{r}}\psi_{ref}(\underline{r},t)}{r^2} + \frac{\hat{\underline{r}}}{rc}\frac{\partial\psi_{ref}(\underline{r},t)}{\partial t}\right]_{ret} \qquad (A5.2.1)$$

We now make the Kirchhoff approximation (Jackson, p280) that the wavefunction incident on Σ is that which would have existed in the absence of Σ. This approximation should be good if multiple reflection is negligible and there is no shadowing, i.e. if the slope of Σ is always small and $h > \max.f(\underline{R})$. If we take ψ to be the velocity potential wavefunction, then the boundary condition at a rigid surface is

$$\text{normal velocity} = \hat{\underline{n}}.\nabla\psi_{incident} + \hat{\underline{n}}.\nabla\psi_{reflected} = 0.$$

It is easy to show that this implies a reflection coefficient of +1. Then we can rewrite (A5.2.1) in terms of the radiated wavefunction as

$$\psi_s(t) = \frac{-1}{4\pi}\int_{\Sigma} d\Sigma\ \hat{\underline{n}}.\left[\frac{-\nabla\psi_{inc}(\underline{r},t)}{r} + \frac{\hat{\underline{r}}\,\psi_{inc}(\underline{r},t)}{r^2} + \frac{\hat{\underline{r}}}{rc}\frac{\partial\psi_{inc}(\underline{r},t)}{\partial t}\right]_{ret}$$

Suppose the radiator at S is a point source with a non-uniform polar diagram $P(\hat{\underline{r}})$. Then

$$\psi_{inc}(\underline{r},t) = P(\hat{\underline{r}})\ \frac{F(t - r/c)}{r}$$

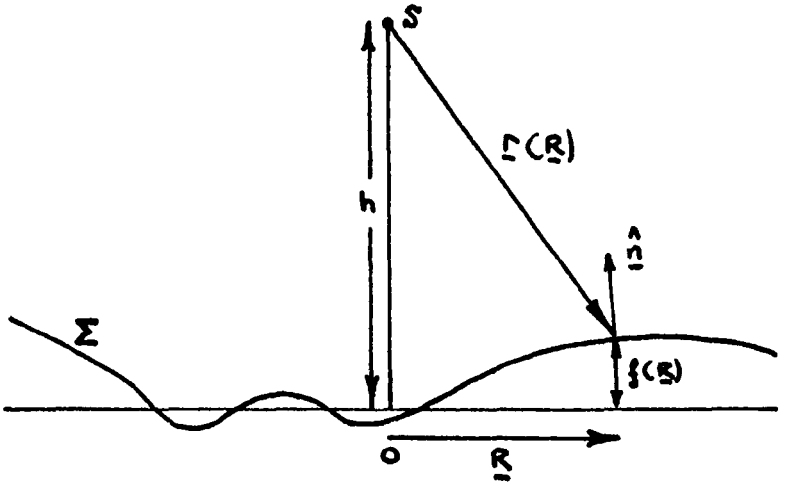


Figure A5.2.1 Geometry for Point Source-Receiver

and $\hat{n}.\nabla\psi_{inc}(\underline{r},t) = \dfrac{F(t - r/c)}{r}\ \underline{n}.\nabla P(\underline{\hat{r}}) + P(\underline{\hat{r}})\ \hat{n}.\hat{r}\left\{\dfrac{-F(t-r/c)}{r^2} - \dfrac{F'(t-r/c)}{rc}\right\}$

giving $\psi_s(t) = \dfrac{-1}{4\pi}\int_\Sigma d\Sigma \left[\dfrac{-F(t-r/c)}{r^2}\ \hat{n}.\nabla P(\hat{r})\right.$

$$\left. + 2\hat{n}.\hat{r}\left\{\frac{F(t-r/c)}{r^3} + \frac{F'(t-r/c)}{r^2c}\right\}P(\hat{r})\right]_{ret}$$

We shall only consider quasimonochromatic pulses F(t) with centre frequency ω, so that

$$|F'(t)| \simeq \omega|F(t)|$$

Then if $\omega/c = k = 2\pi/\lambda >> 1/r$ i.e. $r >> \lambda$, the second term is negligible compared with the third. If also $kP >> \nabla P$, the first term is negligible compared with the third (i.e. if $P(\underline{\hat{r}})$ is slowly varying). Then

$$\psi_s(t) \simeq -\frac{1}{2\pi c}\int_\Sigma d\Sigma\ \hat{n}.\hat{r}\ \left[\frac{F'(t - r/c)}{r^2}\right]_{ret} P(\hat{r})$$

Now following Berry we write

$$\hat{n}.\hat{r} = -\frac{\underline{R}.\nabla_R f(\underline{R}) + h - f(\underline{R})}{r(\underline{R})\sqrt{1+|\nabla_R f(\underline{R})|^2}}$$

and $d\Sigma = \dfrac{d^2\underline{R}}{|n_z|} = \sqrt{1+|\nabla_R f(\underline{R})|^2}\ d^2\underline{R}$

and note that $[F'(t)]_{ret} = F'(t - r/c)$, giving

$$\psi_s(t) \simeq \frac{1}{2\pi c}\int d^2\underline{R}\ \frac{\{\underline{R}.\nabla_R f(\underline{R}) + h - f(\underline{R})\}}{r^3(\underline{R})}\ F'(t - 2r/c)\ P(\hat{r}) \qquad \text{(A5.2.2)}$$

But $r(\underline{R})^2 = \underline{R}^2 + (h - f(\underline{R}))^2$ implies

$$\nabla_R f(\underline{R}) = \frac{r\nabla_R r - \underline{R}}{f(\underline{R}) - h}$$

giving

$$\psi_s(t) \simeq \frac{1}{2\pi c}\int d^2\underline{R}\ \frac{F'(t - 2r(\underline{R})/c)\ P(\hat{r})}{r(\underline{R})\,(h - f(\underline{R}))}\left\{1 - \frac{\underline{R}.\hat{r}}{r(\underline{R})}\right\}$$

Assuming h >> f, we may write the second term as approximately

$$\frac{1}{4\pi}\int d^2\underline{R}\ \frac{P(\hat{r})\ \underline{R}.\nabla_R F}{r^2 h}$$

Integrating by parts and assuming $F(-\infty) = 0$ gives

$$-\frac{1}{4\pi}\int d^2\underline{R}\ F\ \nabla_R\cdot\left(\frac{\underline{R}P(\hat{r})}{r^2 h}\right) \simeq -\frac{1}{2\pi}\int d^2\underline{R}\ \frac{F\ P}{r^2 h}$$

because significant contributions only come from near specular reflections having $\nabla_{\underline{R}} r \simeq 0$. But we have already shown that $|F/r^2| << |F'/r|$, so this term is negligible.

If our receiver also has a response polar diagram $p(\underline{f})$, we can incorporate this into $P(\underline{f})$, and since there is a simple relation between $\underline{f}$ and $\underline{R}$ for given h we can write

$$P(\underline{f})p(\underline{f}) = Q(\underline{R},h)$$

to give finally

$$\psi_s(t) \simeq \frac{1}{2\pi c} \int d^2\underline{R} \; \frac{Q(\underline{R},h)}{r(\underline{R})\,(h-f(\underline{R}))} \; F'(t - 2r(\underline{R})/c) \qquad \text{(A5.2.3)}$$

For a flat surface and a uniform point source this formula gives exactly the right result. We see that the effect of the polar diagrams of the source and receiver are approximately the same as that of a varying attenuation coefficient over the surface, but we shall only use the formula with Q = 1. Note that we have ignored the frequency dependence of the polar diagram: for a quasimonochromatic pulse the centre frequency polar diagram should suffice.

In the next appendix we present the simple Huygens formula used by Landau & Lifschitz to compute the effect of launching a specified wavefront. Then if only paraxial rays are important, so that we can ignore the obliquity factor and set $r(\underline{R}) \simeq h - f(\underline{R})$, (A5.2.3) is approximately

$$\psi_s(t) \simeq \frac{1}{2\pi(c/2)} \int d^2\underline{R} \; \frac{1}{r(\underline{R})} \left\{ \frac{Q(\underline{R},h)F'(t - r/(c/2))}{2r(\underline{R})} \right\}$$

which is approximately the wavefunction produced by launching at t = 0 the wavefront having a height $f(\underline{R})$ and amplitude $Q(\underline{R},h)F'(t)/2r(\underline{R})$, in a medium having wave velocity c/2. This analogy is due to Prof.J.F.Nye.

If the condition $h >> f(\underline{R})$ does not hold, then (A5.2.3) diverges at $h = f(\underline{R})$. In fact, the contribution from such points may be negligibly small so that we should still be able to apply (A5.2.3) with care. However, we should go back to (A5.2.2) which is well defined through $h = f(\underline{R})$, although slightly more complicated. If the surface is such that multiple reflections are negligible and there is no shadowing, despite the fact that it is not everywhere below the transducer, then (A5.2.2) should still be accurate, although (A5.2.3) may not be.

A5.3 Focussing of Pulses

To form a simple picture of the behaviour of dislocations in the neighbourhood of caustics it is necessary to know the effect on the quasimonochromatic pulse wavefunction of the focussing which occurs on the caustic. The simplest possible analysis may be based on the CW analysis given in section 59 of Landau & Lifschitz (1975). Resolve the pulse into its CW components, and consider a wavefront of one of these having wavenumber k. Take some point O on this wavefront as origin, and set up a local coordinate system with x axis along the wavenormal, and y and z axes tangent to the two principal directions of curvature as in fig.A5.3.1. If the two radii of principal curvature are R_1 and R_2, then on the x axis the wave will focus at $x = R_1$ and $x = R_2$, and the equation of the wavefront near the origin is

$$X \simeq y^2/2R_1 + z^2/2R_2$$

The distance from the point (X,y,z) on the wavefront to P ≡ (x,0,0) on the wavenormal is

$$R \simeq x + x_1y^2/2 + x_2z^2/2$$

where $x_1 = 1/x - 1/R_1$ and $x_2 = 1/x - 1/R_2$.

We may compute the approximate continuous wavefunction u_p at P using the Huygen's formula

$$u_p \simeq \int_f \frac{ku}{2\pi iR} e^{ikR} df_n \qquad \text{(A5.3.1)}$$

where u is the wavefunction on the surface f at the point distance R from P, and df_n is the element of area perpendicular to the ray or wavenormal through f at that point (which gives the obliquity factor). We take f to be our wavefront, on which u ≃ constant, giving

$$u_p \simeq \frac{kue^{ikx}}{2\pi ix} \int_{-\infty}^{\infty} dy\, e^{ikx_1y^2/2} \int_{-\infty}^{\infty} dz\, e^{ikx_2z^2/2}$$

$$= \frac{kue^{ikx}}{2\pi ix} \sqrt{\frac{2\pi}{k|x_1|}}\, e^{i(\operatorname{sgn} x_1)\pi/4} \sqrt{\frac{2\pi}{k|x_2|}}\, e^{i(\operatorname{sgn} x_2)\pi/4}$$

$$= \sqrt{\frac{R_1R_2}{|(x-R_1)(x-R_2)|}}\, ue^{ikx}\, e^{-i\theta(x-R_1)\pi/2}\, e^{-i\theta(x-R_2)\pi/2}$$

Now because the frequency dependence of the amplitude has cancelled out, we can easily put in the time variation and resynthesize our

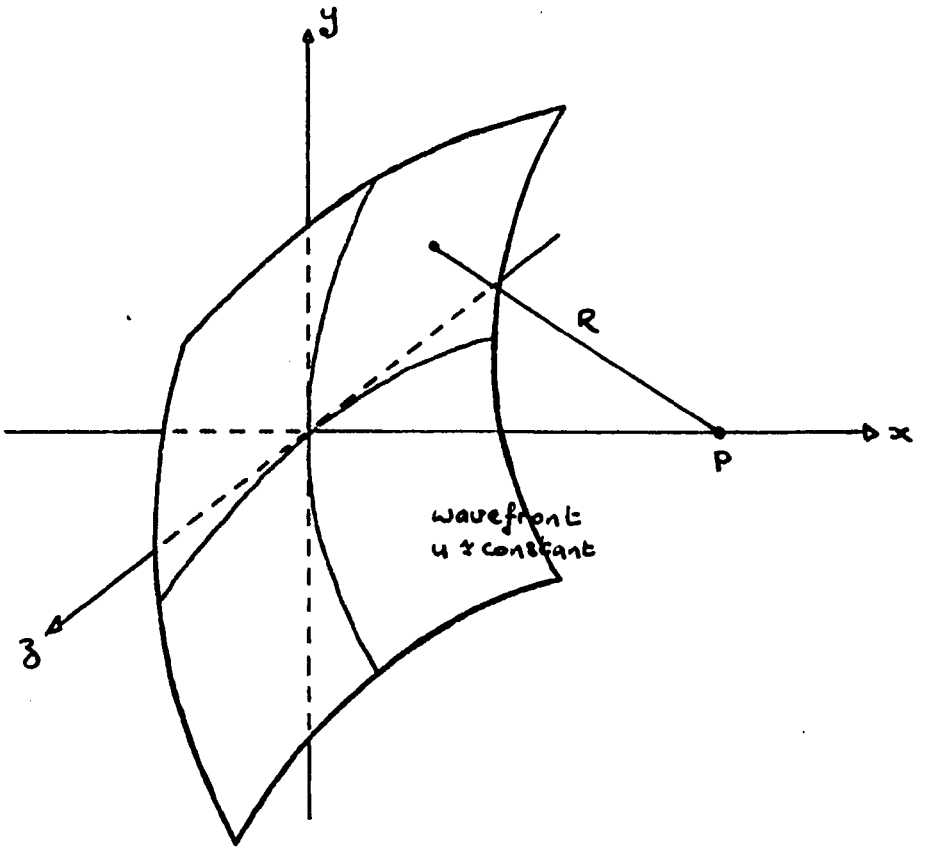


Figure A5.3.1 Converging Wavefront

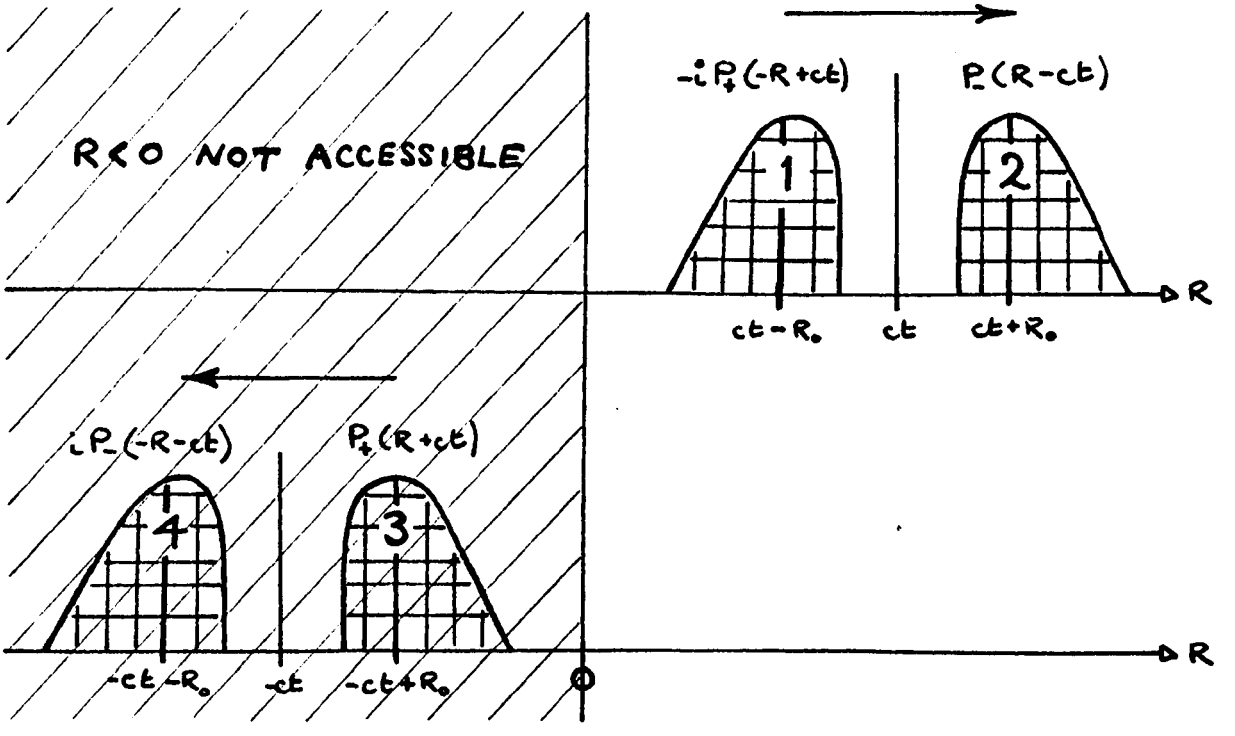


Figure A5.3.2 Schematic Diagram of the Four Pulses

quasimonochromatic pulse to give

$$\Psi_p(x,t) \simeq \sqrt{\frac{R_1R_2}{|(x-R_1)(x-R_2)|}}\ \Psi_0(t-x/c)\ e^{-i\theta(x-R_1)\pi/2}\ e^{-i\theta(x-R_2)\pi/2}$$

where $\Psi_0(t)$ is the wavefunction at the origin. There is an overall amplitude factor which becomes infinite at the two foci, and the phase is retarded by $\pi/2$ through each line focus (or by π through a point focus). Thus the carrier undergoes all the normal focussing effects, while the envelope just "sits on top" of the carrier and is itself completely unaffected. The main effect of passing through a focus is that the carrier is phase shifted relative to the envelope.

The above argument is intuitive and very approximate, and it would be nice to solve the problem more rigorously. We shall therefore analyse the passage of a quasimonochromatic pulse through a perfect line focus, by solving the wave equation with circular symmetry. We require our wavefunction to exist throughout the $\underline{R}$ plane, and in particular at the origin, therefore we may write it as an angular spectrum of plane waves

$$\Psi(\underline{R},t) = \int d^2\underline{K}\ G(\underline{K})\ e^{i(\underline{K}.\underline{R} - cKt)}$$

For circular symmetry this reduces to

$$\Psi(\underline{R},t) = \int_0^\infty K\ dK\ G(K) \int_0^{2\pi} d\theta\ e^{i(KR\cos\theta - Kct)}$$

$$= 2\pi \int_0^\infty K\ dK\ G(K)\ J_0(KR)\ e^{-iKct}$$

where $J_0(KR)$ is a Bessel function. The fact that the domain of $\Psi(\underline{R},t)$ includes $R = 0$ precludes the appearance of $Y_0(KR)$, the singular Bessel function. Now if we try to set up a pulse by imposing a boundary condition at some radius we shall have the complication of multiple reflections, although this is the only way to create initially a purely converging pulse. Instead we shall impose an initial condition to create a disturbance fixed in space, and then let it go at time $t = 0$. We would expect the disturbance to spread initially as a converging pulse plus a diverging pulse, and we wish to study the converging pulse.

The initial wavefunction is

$$\Psi(R,0) = \int_0^\infty K\ dK\ 2\pi\ G(K)\ J_0(KR).$$

This is a Hankel transform (see Watson (1944)) which we can invert if

$$\int_0^\infty \Psi(R,0)\ \sqrt{R}\ dR$$

exists and is absolutely convergent. This will be the case, because $\Psi(R,0)$ will be peaked around $R = R_0 >> 0$. Then

$$2\pi\, G(K) = \int_0^\infty R'\, dR'\, \Psi(R',0)\, J_0(KR')$$

so that

$$\Psi(R,t) = \int_0^\infty R'\, dR'\, \Psi(R',0)\, \chi(R,R',t) \qquad \text{(A5.3.2)}$$

where

$$\chi(R,R',t) = \int_0^\infty K\, dK\, J_0(KR)\, J_0(KR')\, e^{-iKct}$$

is the propagator. This integral cannot usefully be evaluated in terms of simple functions, but we are only interested in the wavefunction at large radii, and our initial wavefunction will be such that only large K are important. Therefore we can use the asymptotic forms for the Bessel functions, viz.

$$J_0(x) \sim \left(\frac{2}{\pi x}\right)^{\frac{1}{2}} \cos(x - \pi/4)$$

so that $\chi(R,R',t) \sim \frac{1}{2\pi} \left(\frac{1}{RR'}\right)^{\frac{1}{2}} \int_0^\infty dK \left\{ e^{iK(R+R'-ct)-i\pi/2} + e^{iK(R-R'-ct)} \right.$

$$\left. + e^{iK(-R+R'-ct)} + e^{iK(-R-R'-ct)+i\pi/2} \right\}$$

Let us define

$$D(x) = \frac{1}{\pi} \int_0^\infty dK\, e^{iKx} = \delta(x) + \frac{i}{\pi x}$$

in the sense of generalized functions (see Lighthill (1958)). Then

$$\Psi(R,t) \sim \frac{1}{2R^{\frac{1}{2}}} \int_0^\infty dR'\, F(R') \left\{ -iD(R+R'-ct) + D(R-R'-ct) \right.$$

$$\left. + D(-R+R'-ct) + iD(-R-R'-ct) \right\}$$

where $F(R') = R'^{\frac{1}{2}}\, \Psi(R',0)$, which on changing the variable of integration in each term becomes

$$\Psi(R,t) \sim \frac{1}{2R^{\frac{1}{2}}} \left\{ -i \int_{R-ct}^\infty dx F(x-R+ct) D(x) + \int_{-\infty}^{R-ct} dx F(-x+R-ct) D(x) \right.$$

$$\left. + \int_{-R-ct}^\infty dx F(x+R+ct) D(x) + i \int_{-\infty}^{-R-ct} dx F(-x-R-ct) D(x) \right\}$$

$$= \frac{1}{2R^{\frac{1}{2}}} \int_{-\infty}^\infty dx \left\{ -i\theta(x-R+ct) F(x-R+ct) + \theta(-x+R-ct) F(-x+R-ct) \right.$$

$$\left. + \theta(x+R+ct) F(x+R+ct) + i\theta(-x-R-ct) F(-x-R-ct) \right\} D(x)$$

Now $D(x)$ is sharply peaked about $x = 0$, so we can set $x = 0$ in the θ-functions and take them outside the integrals without incurring much error unless the arguments are close to 0. If we also define

$$P_{\pm}(y) = \int_{-\infty}^{\infty} dx\ F(\pm x + y)\ D(x)$$

then

$$\Psi(R,t) \sim \frac{1}{2R^{\frac{1}{2}}} \{-i\theta(-R+ct)P_{+}(-R+ct) + \theta(R-ct)P_{-}(R-ct)$$
$$+ \theta(R+ct)P_{+}(R+ct) + i\theta(-R-ct)P_{-}(-R-ct)\}$$

asymptotically as $R \to \infty$. Terms (1) and (2) represent outgoing pulses; (1) for $R < ct$ and (2) for $R > ct$. Similarly terms (3) and (4) represent ingoing pulses; (3) for $R > -ct$ and (4) for $R < -ct$. Thus the form of the solution changes at $R = \pm ct$. We represent this schematically in fig.A5.3.2 where we assume $P_{\pm}(R)$ is peaked non-symmetrically about R_0. In fact, because of the peaked nature of the pulses, the θ-functions do not have any significance. For small t the initial disturbance splits up into pulses (2) and (3), (2) moving out and (3) moving in, as anticipated. In the limit $t = 0$, we have

$$\Psi(R,0) \sim \frac{1}{2R^{\frac{1}{2}}} \{P_{-}(R) + P_{+}(R)\} = \frac{F(R)}{R^{\frac{1}{2}}}$$

as we originally specified. Pulses (1) and (2) move "rigidly" out, and pulses (3) and (4) move "rigidly" in, as t increases. Pulse (2) just runs off to infinity, and pulse (4) is never seen for $t > 0$. Pulse (3) passes through the focus at $t \simeq R_0$, and reappears as pulse (1). This is the process that we are interested in. Let us write

$$P_{+}(R) = Q(R - R_0)$$

where $Q(R)$ is peaked about $R = 0$. Then we start with the pulse

$$\Psi(R,t) \simeq \frac{1}{2R^{\frac{1}{2}}} Q(R + ct - R_0) \qquad \text{(A5.3.3)}$$

for $t \ll R_0/c$, which focusses at $t \simeq R_0/c$ and reappears as

$$\Psi(R,t) \simeq \frac{-i}{2R^{\frac{1}{2}}} Q(-R + ct - R_0) \qquad \text{(A5.3.4)}$$

for $t \gg R_0/c$. Put $t = 2R_0/c - T$ in (A5.3.4) to give

$$\Psi(R,2R_0/c-T) \simeq \frac{-i}{2R^{\frac{1}{2}}} Q(-\{R + cT - R_0\})$$

which is exactly -i times the converging pulse but "back to front" and diverging. Asymptotically the pulse appears to have travelled straight through the focus with constant velocity c. It suffers no time delay or distortion, and the only effect of the focus is to multiply the

pulse by -i which retards the phase by $\pi/2$, as we found by our first crude argument.

The only approximations made here are firstly to use the asymptotic forms of the Bessel functions, which is justified if our initial pulse is quasimonochromatic and peaked sufficiently sharply about some large radius R_0, and we do not look at the pulses too close to the focus. Secondly we have replaced x in the θ-functions by 0, which is justified well away from $R = \pm ct$. But because of the peaked nature of the pulses this region is not of interest (see fig. A5.3.2). Therefore our deductions should be accurate, and our crude intuitive argument is vindicated.

CHAPTER 6

THE SWALLOWTAIL DIFFRACTION PATTERN

6.1 The Canonical Integral and Caustic

We shall study the integral

$$S(W_1,W_2,W_3) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\ e^{i\phi(x,W_1,W_2,W_3)} \tag{6.1}$$

where $\phi(x,W_1,W_2,W_3) = x^5/5 + W_3x^3/3 + W_2x^2/2 + W_1x$ (6.2)

The transformation $x \to -x$ shows that $S(W_1,-W_2,W_3) = S^*(W_1,W_2,W_3)$: S has no other symmetry. Hence the amplitude, and therefore the caustic, is symmetric in W_2, and the phase is antisymmetric.

The first step in any study of these integrals is to find the bifurcation set B, which is given by

$$\left.\begin{aligned} \partial\phi/\partial x &= x^4 + W_3x^2 + W_2x + W_1 = 0 \qquad (6.3a)\\ \partial^2\phi/\partial x^2 &= 4x^3 + 2W_3x + W_2 = 0 \qquad (6.3b) \end{aligned}\right\}$$

These equations are non-linear in x, although linear in $\underline{W}$, and sufficiently complicated that we cannot eliminate x to give the explicit equation of B as we did for the cusp. But we can solve explicitly for any two components of $\underline{W}$ in terms of x and the third component of $\underline{W}$ to give B parametrically. In particular, we can find the parametric equations of any plane section through B, parametrised by the state variable x. This provides a simple general method for plotting the bifurcation sets of all the elementary catastrophes.

The planes we are most interested in are W_3 = constant, in which

$$\left.\begin{aligned} W_1 &= x^2(3x^2 + W_3)\\ W_2 &= -2x(2x^2 + W_3) \end{aligned}\right\} \tag{6.4}$$

For a caustic, x must be real => $x^2 \geqslant 0$. For $W_3 = 0$, the explicit equation is

$$W_1 = 3(W_2/4)^{4/3}$$

(see fig.6.1b) which has a bend of infinite curvature at the origin (the swallowtail singularity). As W_3 increases the curve broadens out and becomes everywhere regular, tending toward the very flat parabola

$$W_1 = W_2^2/4W_3$$

as $W_3 \to \infty$ (see fig.6.1a).

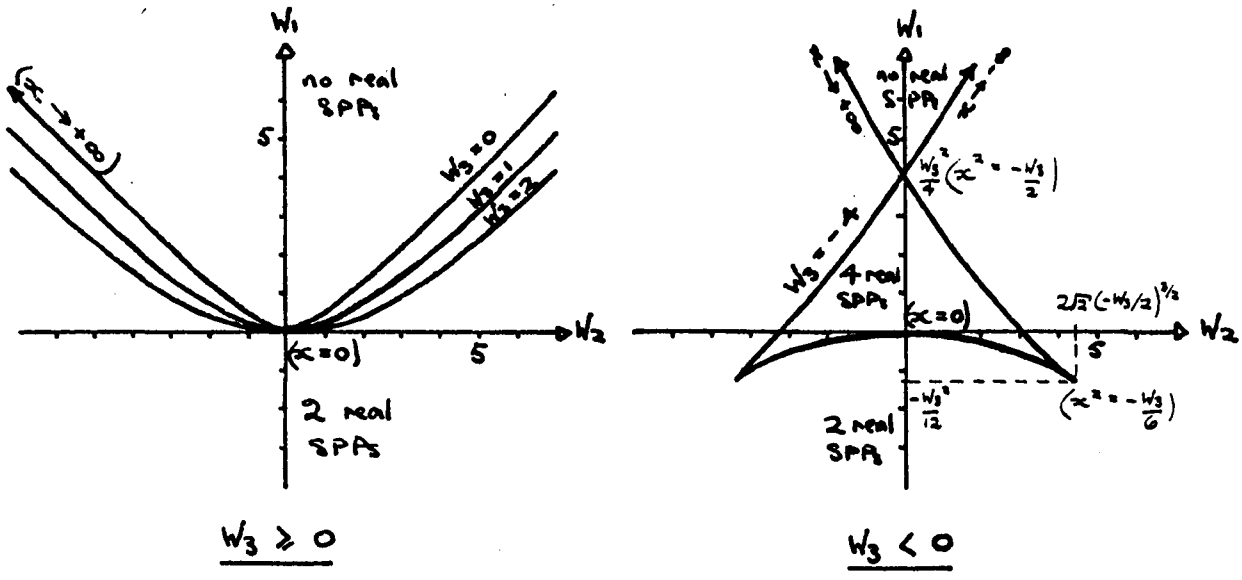


Figure 6.1 W_3 = constant sections of Swallowtail Caustic

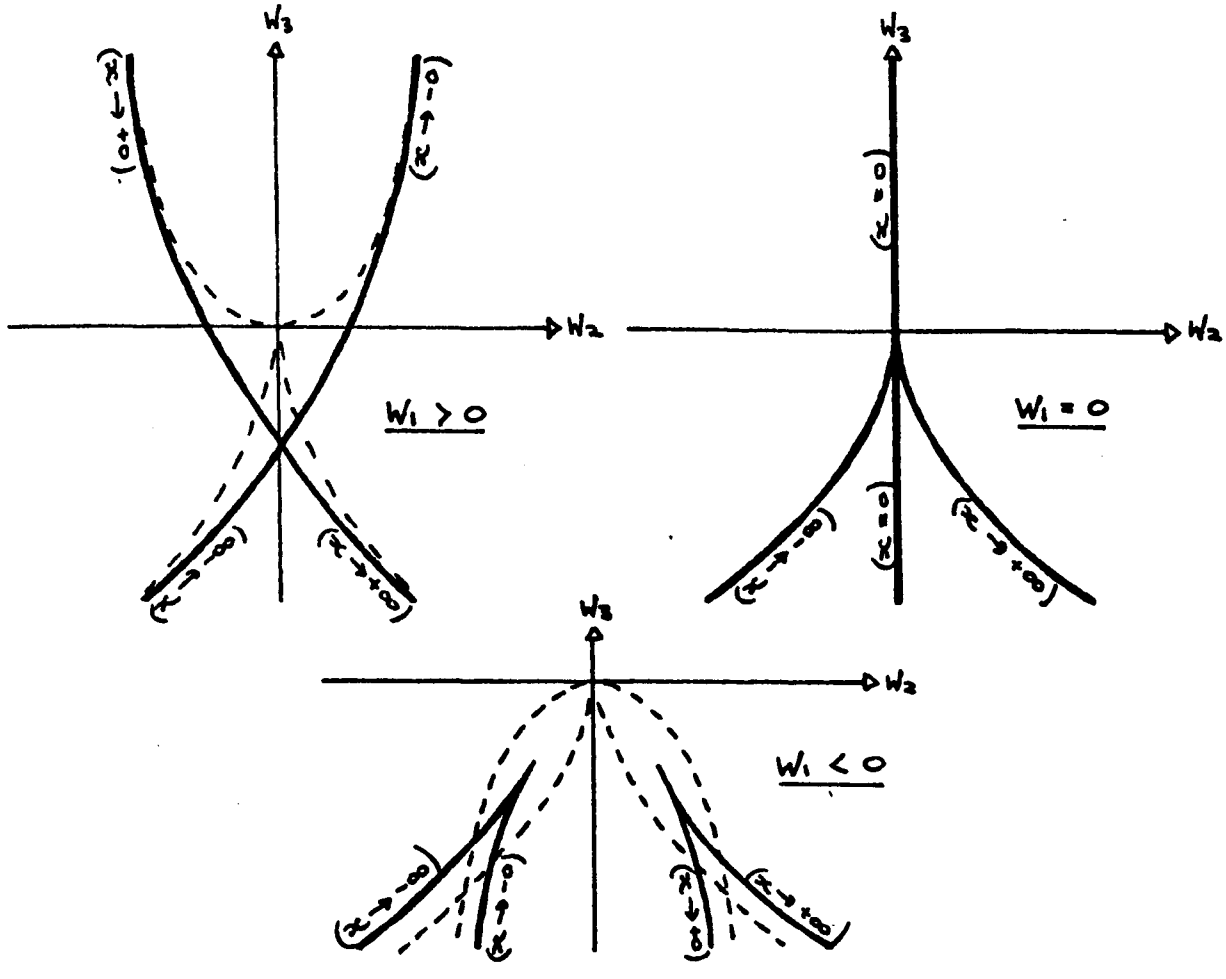


Figure 6.2 W_1 = constant sections of Swallowtail Caustic

Notice that

$$\partial W_1/\partial x = 2x(6x^2 + W_3)$$

$$\partial W_2/\partial x = -2(6x^2 + W_3)$$

showing that for $W_3 < 0$ there are cusp singularities at $x = \pm\sqrt{-W_3/6}$, and there is also a double point at $x = \pm\sqrt{-W_3/2}$. The details are summarized in fig.6.1, and we see that B includes the whole W_3 axis. For $W_3 > 0$, B is a regular surface, but for $W_3 < 0$ there are two cuspidal lines and a line of self-intersection running from the principal singularity at $\underline{W} = 0$.

By similar means we sketch the sections W_1 = constant in fig.6.2 (where the asymptotic forms of a parabola and a semi-cubical parabola are shown dashed), and W_2 = constant in fig.6.3. In fig.6.4 we sketch the whole bifurcation set: upside down it resembles the tail of a swallow! We discuss the details of plotting the W_3 sections in the appendix.

We can solve the ray equation (6.3a) exactly when $W_2 = 0$, and then by continuity we see that there are no real rays in "the body of the swallow", which we therefore call the dark region. Outside the swallow there are 2 real rays, and inside the tail there are 4 real rays, so this region should be the brightest. The folding of the catastrophe manifold which produces these numbers of sheets in each region is shown in the pictures by Woodcock & Poston (1974). In the regions where there are real SPPs we can apply our standard integration method, the details of which are described in the appendix. Just inside the dark region we can still integrate over the region where $\phi(x)$ is most slowly varying. But as we move further into the dark region away from the caustic, the minimum value of $\phi_1(x)$ increases, and the integrand eventually oscillates too rapidly for a quadrature routine to handle. We need a new numerical method based on the method of steepest descent, which is the generalization of the method of stationary phase to complex SPPs.

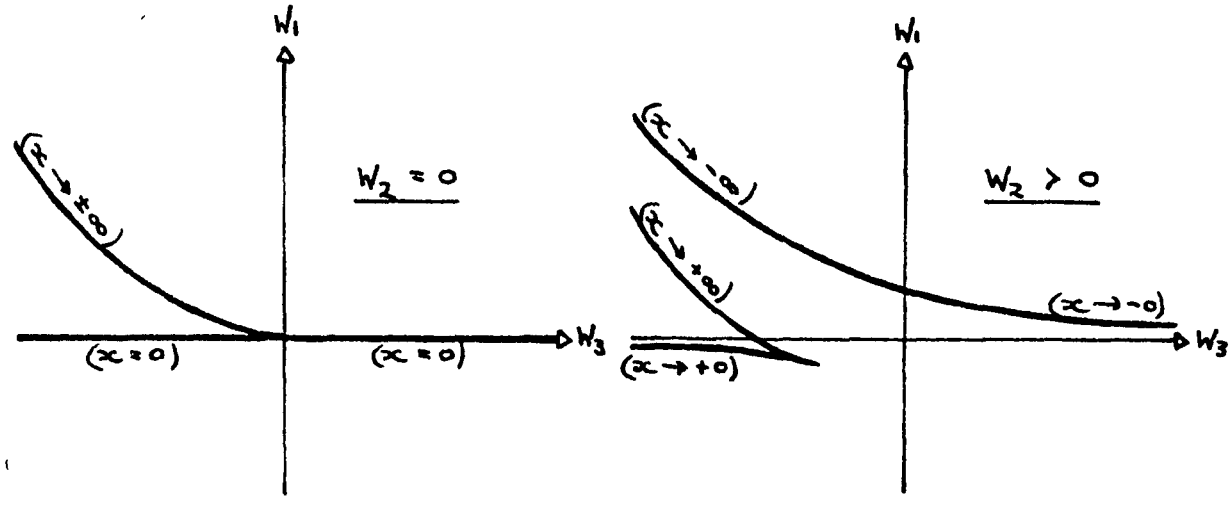


Figure 6.3 W_2 = constant sections of Swallowtail Caustic

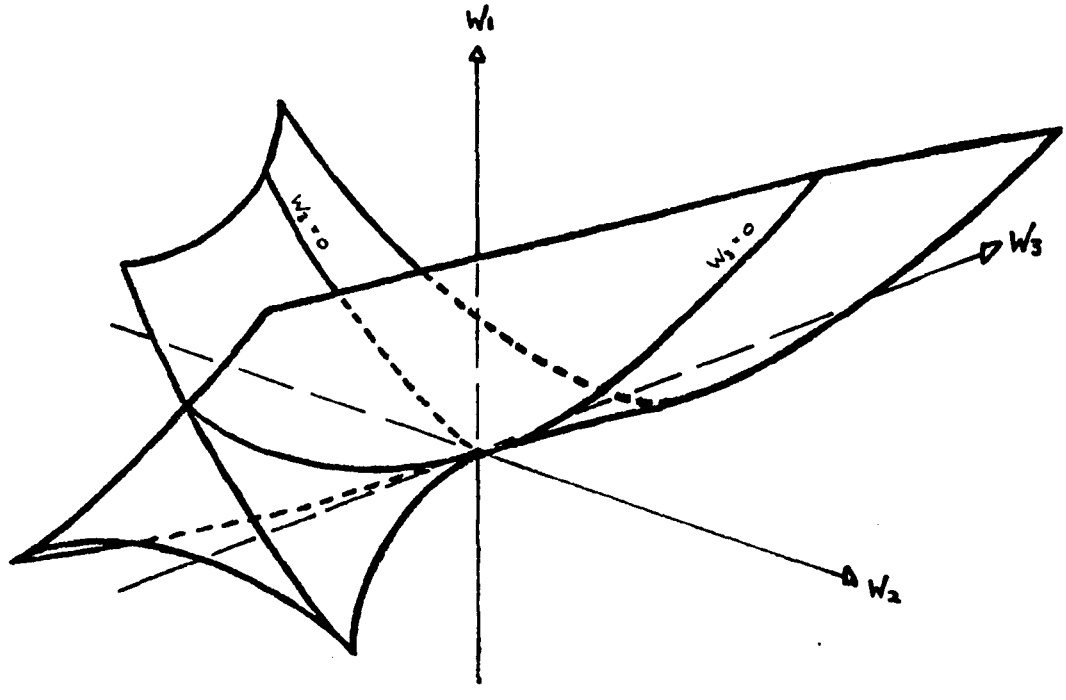


Figure 6.4 The Swallowtail Caustic

6.2 The Dark Region

In the dark region the SPPs are all complex. The 4 roots of (6.3a) occur in 2 complex conjugate pairs, giving 2 saddles of $\phi(x)$ in the upper half plane plus their mirror images in the lower half plane. If two of these complex SPPs coalesce, (6.3b) is also satisfied. For complex x and real $\underline{W}$, (6.3) represents 4 equations in 5 unknowns which we can solve in terms of W_3. The only complex solution for x is

$$x = \pm i\sqrt{W_3/2}, \quad W_1 = (W_3/2)^2, \quad W_2 = 0 \quad \text{for } W_3 \geq 0.$$

The complex SPPs coalesce into two double SPPs along a half line in control space running from the origin to infinity. This is the reflection in the plane $W_3 = 0$ of the line of self-intersection of the caustic: it is called a "complex whisker" (see Poston & Stewart (1976)). It is interesting that real SPPs coalesce on a <u>surface</u> in control space, while complex SPPs only coalesce along a <u>line</u> in control space, because an extra condition has to be satisfied. We shall see later that the complex whisker is <u>not</u> in any sense a caustic.

Initially we will consider $W_3 \geq 0$, the "positive side" of the swallowtail. For $W_2 = 0$ we can solve (6.3a) exactly for the complex saddles β as a function of W_1, and we sketch their trajectories in fig.6.5a. The identity of the saddles where they coalesce is ambiguous, but for small $W_2 > 0$ and small $\delta W_1 = W_1 - (W_3/2)^2$ we can expand ϕ_1 about the complex double saddle to give

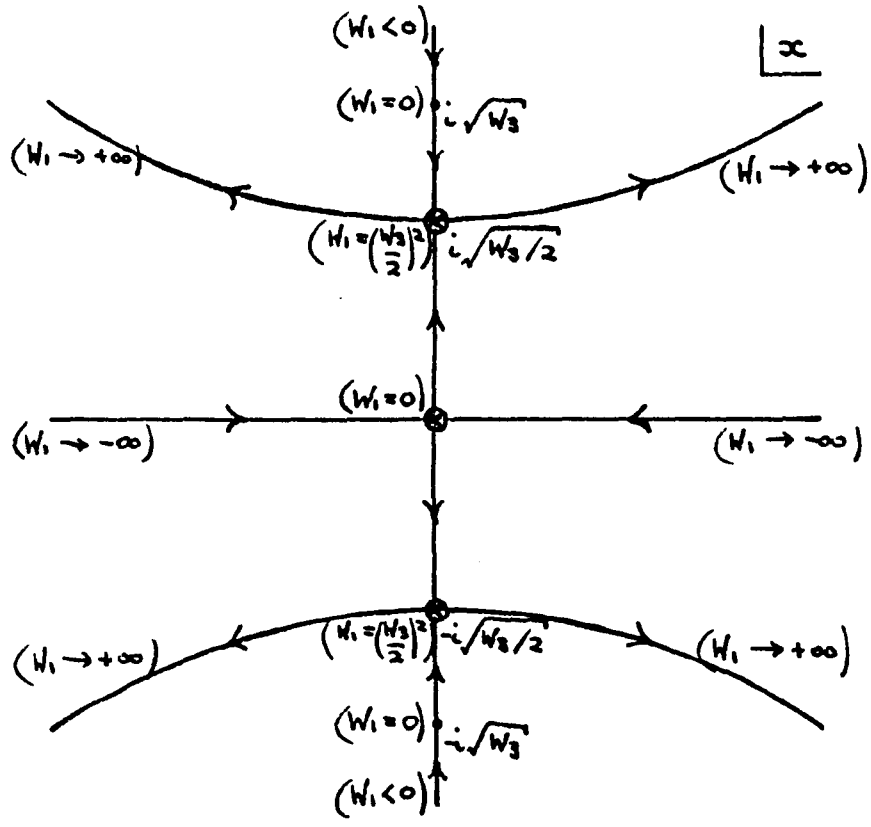
$$\beta_{2/1} \simeq i\sqrt{W_3/2} \; +/- \left(\frac{\delta W_1 + i\sqrt{W_3/2}\, W_2}{2W_3}\right)^{\frac{1}{2}}$$

Note that the complex saddles satisfy $\beta_3 = \beta_1^*$ and $\beta_4 = \beta_2^*$, and because (6.3a) has no cubic term $\Sigma\beta$'s = 0 $\Rightarrow$ Re β_1 = -Re β_2. Expanding ϕ_1 about the real double saddle gives

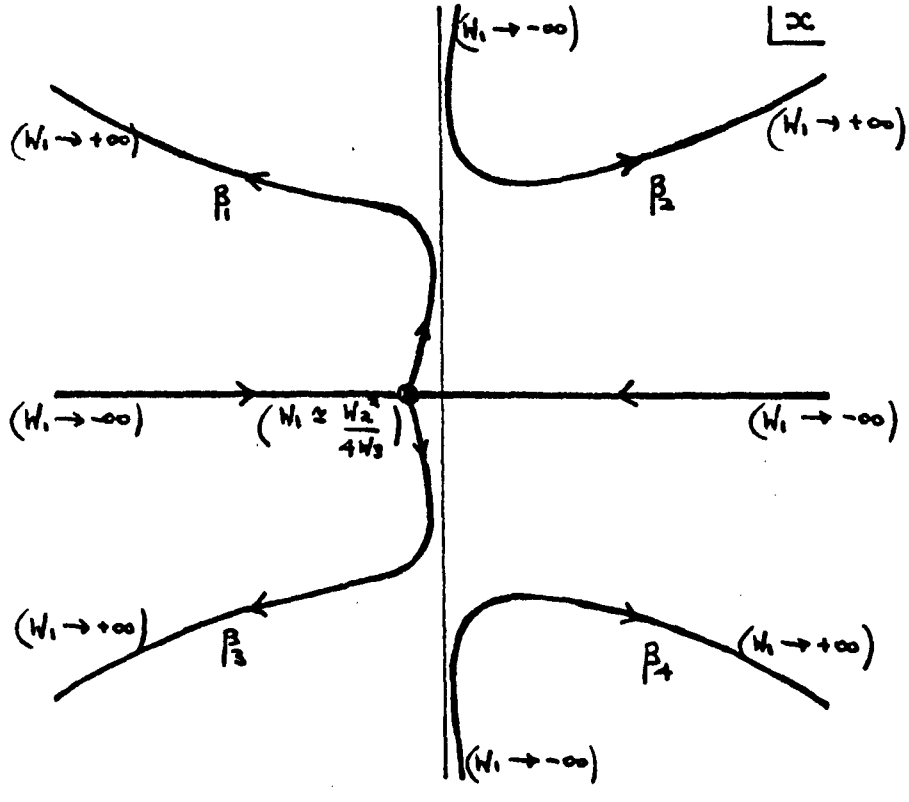
$$\beta \simeq \frac{-W_2 \pm \sqrt{W_2^2 - 4W_1W_3}}{2W_3}$$

We have removed the complex double root (and hence the ambiguity there) only, as sketched in fig.6.5b.

It is easy to deduce the local form of $\phi(x)$ about its saddles for $W_2 = 0$, and from this plus the asymptotic behaviour to deduce the overall topography of $|e^{i\phi(x)}| = e^{-\mathrm{Im}\,\phi(x)}$. Then by continuity we can deduce the topography for $W_2 > 0$. To apply the method of steepest descents we distort the integration contour in the finite x plane to lie along lines of steepest descent from saddles of $|e^{i\phi}|$, using whichever of the 4



(a) $W_2 = 0$ 1 real double saddle & 2 complex double saddles



(b) Small $W_2 > 0$ Only real double saddle remains

Figure 6.5 Trajectories of Saddles of $\phi(x)$

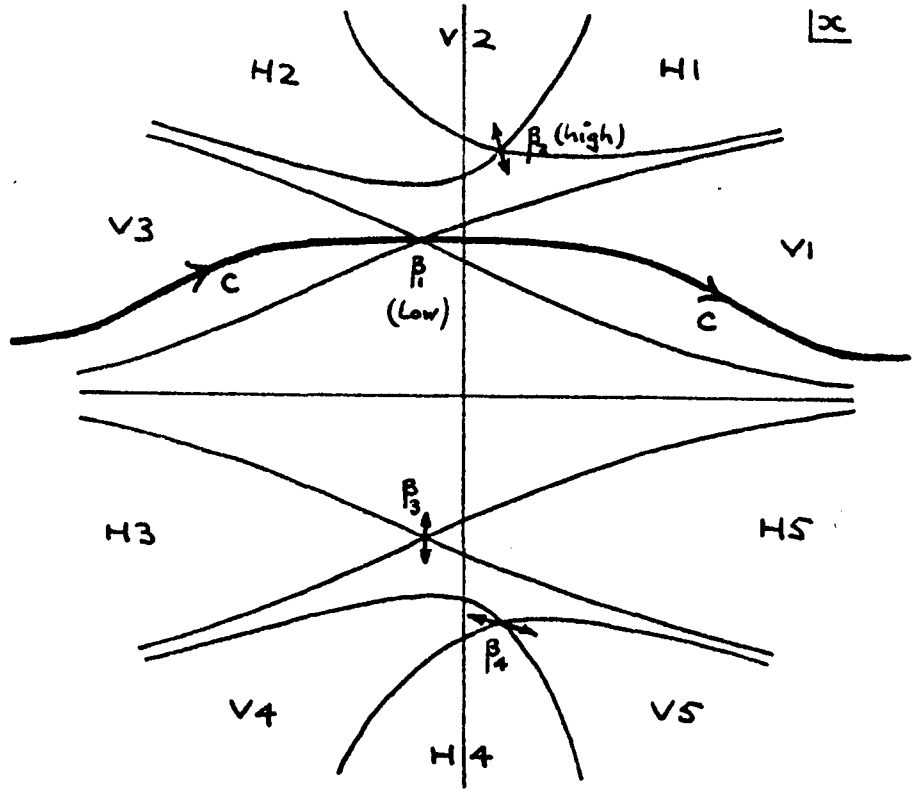
→ denotes motion as W_1 increases

⊗ denotes double saddle

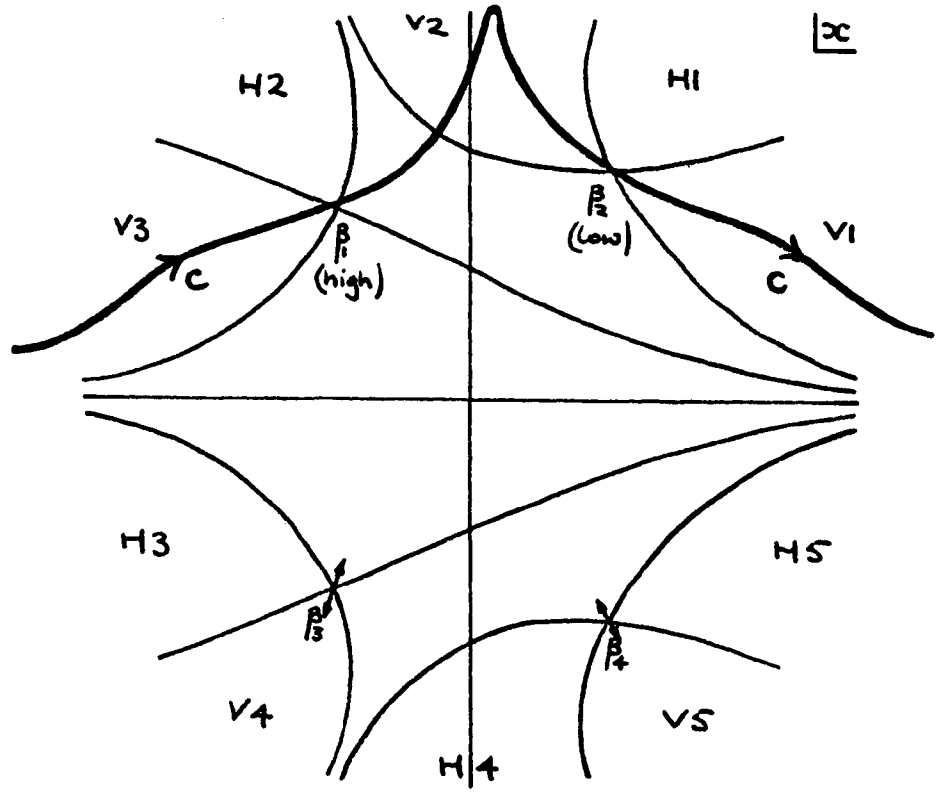
saddles prove necessary. Then the saddles used will give local maxima of $|e^{i\phi}|$, so that their neighbourhoods give the dominant contribution to the integral. We must not, of course, distort a contour through any singularity. Then the topography, and necessary contours, for $W_2 > 0$ are shown in figs.6.6a & b for $0 < W_1 < (W_3/2)^2$ and $W_1 > (W_3/2)^2$ respectively.

We notice that the saddles in the lower half plane do not contribute at all. For small W_1 only β_1 contributes, giving only one complex ray and no chance of any interference. For large W_1, β_1 and β_2 contribute giving 2 complex rays which will interfere and may produce nulls. There must be a surface $W_1(W_2,W_3)$ on which the number of complex rays changes. Obviously for $W_2 = 0$, this will occur at the double saddle, so the complex whisker is contained in this surface. In a plane W_3 = constant we are dealing with an integrand which is parametrised by two variables W_1 and W_2, and locally has two saddles which may coalesce. This is exactly analogous to the integrand of the Airy integral function Ai(z) for <u>complex</u> z (see Budden (1961)), so we shall borrow the terminology applied originally to the asymptotic approximation of the Airy function. We shall call the surface on which the number of contributing complex rays changes (by one) the "Stokes bifurcation set", SB. Since the contribution from β_2 comes in discontinuously, it must be subdominant i.e. much less than that due to β_1, so that the representation of the integral is approximately continuous. Then $|e^{i\phi(\beta_2)}/e^{i\phi(\beta_1)}|$ will be a minimum on SB, and the line of steepest descents through β_1 (which satisfies Re $\phi(\beta_1)$ = constant) runs down to β_2 where it turns through 90^o. This it may do equally well in either direction, and it is this indeterminacy which allows the contour to change over, as shown in fig.6.7a.

However, near to the real bifurcation set B (see fig.6.6a), β_2 is dominant, not subdominant. Therefore, between B and SB there must be a surface on which β_1 and β_2 exchange dominance, which we call the "Anti-Stokes set", AS. Here $|e^{i\phi(\beta_2)}| = |e^{i\phi(\beta_1)}| \Rightarrow$ Im $\phi(\beta_2)$ = Im $\phi(\beta_1)$ and the two saddles are at the same height. For $W_2 = 0$, this equality holds for all $W_1 \geq (W_3/2)^2$, (see fig.6.6b) hence the plane $W_2 = 0$ above the complex whisker is part (AS') of AS. This is because the transformation $W_2 \rightarrow -W_2$ is equivalent to Re(x) $\rightarrow$ -Re(x), so the two saddles must swap dominance. Similarly the plane $W_2 = 0$ below the complex whisker is formally part (SB') of SB (see fig.6.6a).



(a) $W_2 > 0, \quad 0 < W_1 < (W_3/2)^2$



(b) $W_2 > 0, \quad W_1 > (W_3/2)^2$

Figure 6.6 Topography of $|e^{i\phi(x)}|$ and Integration Contour C showing level lines and lines of steepest descent through saddles. H denotes hills, V denotes valleys.

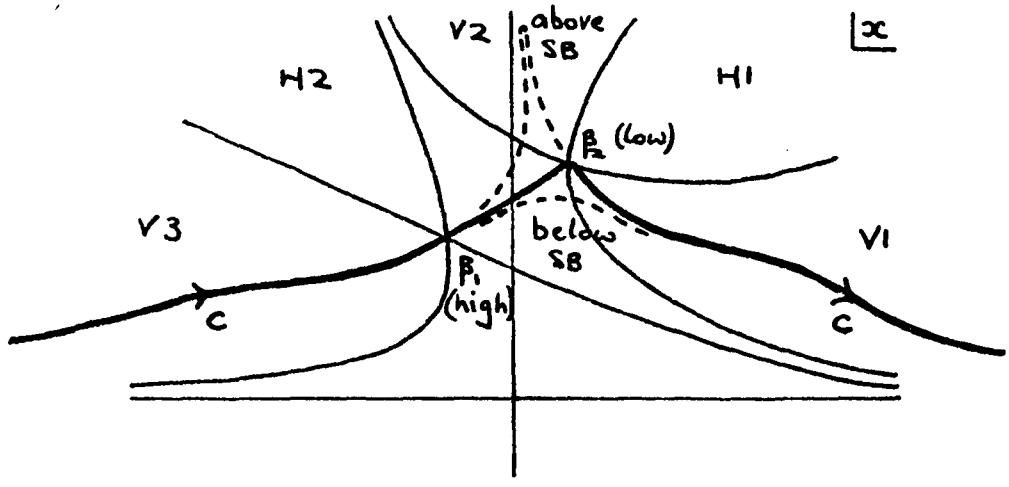


Figure 6.7a Integration Contour C on Stokes Bifurcation Set SB

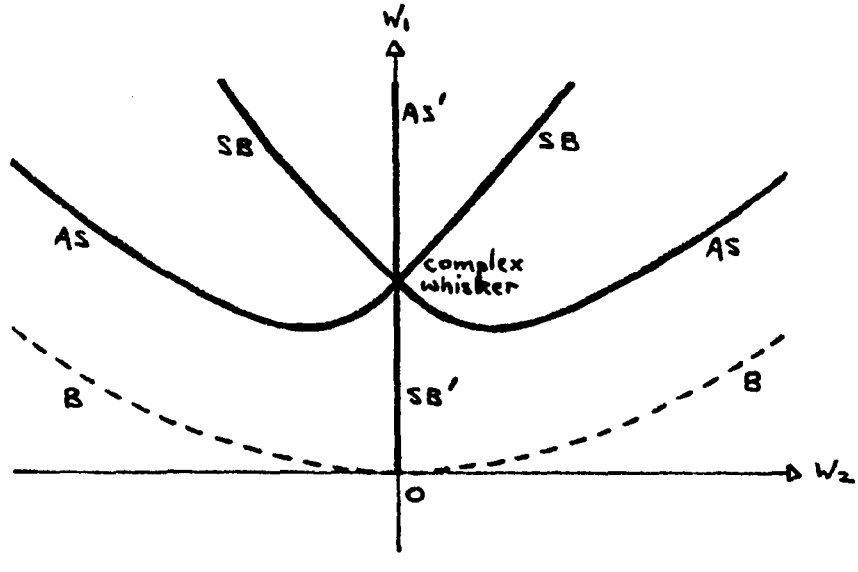


Figure 6.7b The Neighbourhood of the Double Saddle

We can expand ϕ about the complex whisker, using the local forms of β_1 and β_2 derived earlier, to give

$$\Delta\phi \equiv \phi(\beta_2) - \phi(\beta_1) \simeq 2(\delta W_1 + i\sqrt{W_3/2}\, W_2)\left(\frac{\delta W_1 + i\sqrt{W_3/2}\, W_2}{2W_3}\right)^{\frac{1}{2}}$$

$$\propto (\delta W_1 + i\sqrt{W_3/2}\, W_2)^{3/2}$$

This shows that locally SB and AS, given respectively by Re $\Delta\phi = 0$ and Im $\Delta\phi = 0$, have the joint equations

$$W_2 = 0, \quad \delta W_1 = \pm\sqrt{W_3/6}\, W_2$$

as illustrated in fig.6.7b for $W_3 = 6$.

For $W_3 < 0$ there are no complex double saddles and always 2 complex rays above B. Therefore SB and AS must merge into B at $W_3 = 0$.

6.3 Integration by Steepest Descent

The lines of steepest descent have equation Re $\phi(x)$ = constant, therefore if Re $\phi(\beta_1)$ = Re $\phi(\beta_2)$ then the two saddles share a common line of steepest descent. Then if Re $(\phi(\beta_1) - \phi(\beta_2))$ changes sign we know that we have crossed SB, which is important computationally. Crossing AS is not. We evaluate the contribution to S of the contributing saddles by the uniform approximation methods of Chester et al.(1957). About each isolated saddle at β, we map x onto a new variable u defined by

$$-\tfrac{1}{2}u^2 = i\{\phi(x) - \phi(\beta)\},$$

so that the contribution of the saddle at β is approximately

$$S_\beta \simeq \frac{1}{\sqrt{2\pi}}\, e^{i\phi(\beta)} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} \frac{dx}{du}\, du$$

where $dx/du = iu/\phi_1(x)$. There exists an expansion

$$dx/du = \sum_{m=0}^{\infty} c_m u^m$$

with radius of convergence not greater than the distance to the nearest saddle point, where dx/du diverges. Successively differentiating and setting u = 0, x = β gives

$$\frac{dx}{du} = \sqrt{\frac{i}{\phi_2}} \left\{ 1 + \frac{1}{4\phi_2^2}\left(\frac{5\phi_3^2}{3\phi_2^2} - \phi_4\right)\frac{u^2}{2} + O(u^4) \right\}\Bigg|_{x=\beta}$$

+ odd terms which do not contribute.

We assume $\phi_2(\beta)$ is large enough that the rest of this series is negligible compared with the first term. Now the x contour is always traversed from left to right, implying

$$|\arg dx/du| < \pi/2 \implies \text{Re } \sqrt{i/\phi_2(\beta)} > 0$$

which fixes the ambiguity in the square root. Then

$$S_\beta \simeq \frac{e^{i\phi(\beta)}}{\sqrt{2\pi}} \sqrt{\frac{i}{\phi_2(\beta)}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\, du = \sqrt{\frac{i}{\phi_2(\beta)}}\; e^{i\phi(\beta)}$$

if the nearest saddle is sufficiently far away. From a table of the error function, we find that if we only integrate in u between ±2.58, instead of ±∞, we introduce 1% error. This is approximately equivalent to integrating out a distance $2.58/\sqrt{\phi_2(\beta)}$ along the x contour either side of β. We require that our integral has converged to sufficient accuracy before we hit another saddle point, therefore we require

$$2.58/\sqrt{\phi_2(\beta)} < |\beta_1 - \beta_2| \qquad (6.5)$$

Note that this is only an order of magnitude estimate: the accuracy can be increased by increasing the figure of 2.58 if the results suggest that this is necessary.

When (6.5) fails to hold, the two saddles are so close together that they can no longer be treated independently, and $\phi(x)$ is locally cubic, not quadratic. Therefore in the neighbourhood of the pair of saddles we map x onto u by setting

$$\phi(x,\underline{W}) = -(u^3/3 + Z(\underline{W})u) + A(\underline{W}) \tag{6.6}$$

(The minus sign is chosen to preserve the topography of the integrand, and therefore the contour, see fig.6.8). We require (6.6) to be a uniformly regular (1::1) transformation => $dx/du \neq 0$ or ∞, where

$$\phi_1(x)\ dx/du = -(u^2 + Z) \tag{6.7}$$

Then the RHS must vanish when the LHS vanishes at $x = \beta_1$ or β_2. Let us choose

$$\begin{aligned} u &= +iZ^{\frac{1}{2}} \quad \text{at} \quad x = \beta_2 \\ u &= -iZ^{\frac{1}{2}} \quad \text{at} \quad x = \beta_1 \end{aligned} \tag{6.8}$$

Substituting in (6.6) gives

$$\begin{aligned} A(\underline{W}) &= \tfrac{1}{2}\{\phi(\beta_2) + \phi(\beta_1)\} \\ Z^{\frac{1}{2}}(\underline{W}) &= \left[\{\phi(\beta_2) - \phi(\beta_1)\}3i/4\right]^{1/3} \end{aligned} \tag{6.9}$$

Very close to the complex whisker, the local approximation of $\phi(\beta_2)-\phi(\beta_1)$ in section 6.2 gives

$$Z(\underline{W}) \simeq 2(-9)^{1/3}(\delta W_1 + i\sqrt{W_3/2}\ W_2)$$

which is just the $\underline{W}$ plane about the whisker inverted in the origin and squashed. Since we have fixed the contour as in fig.6.8 we must choose the cube root to give the saddles in u in the right places relative to the contour (see later).

We can expand dx/du in the form

$$dx/du = \sum_{m=0}^{\infty}(p_m + uq_m)(u^2 + Z)^m \tag{6.10}$$

Differentiating (6.7) and substituting (6.8) gives

$$\left.\frac{dx}{du}\right|_{\beta_2} = \sqrt{\frac{-2iZ^{\frac{1}{2}}}{\phi_2(\beta_2)}} = R_2 \quad \text{and} \quad \left.\frac{dx}{du}\right|_{\beta_1} = \sqrt{\frac{+2iZ^{\frac{1}{2}}}{\phi_2(\beta_1)}} = R_1$$

The mapping $x \to u$ has been chosen so that rotations in the complex plane are small, therefore for both these square roots we require

$$|\arg R| < \pi/2 \quad => \quad \mathrm{Re}\ R > 0.$$

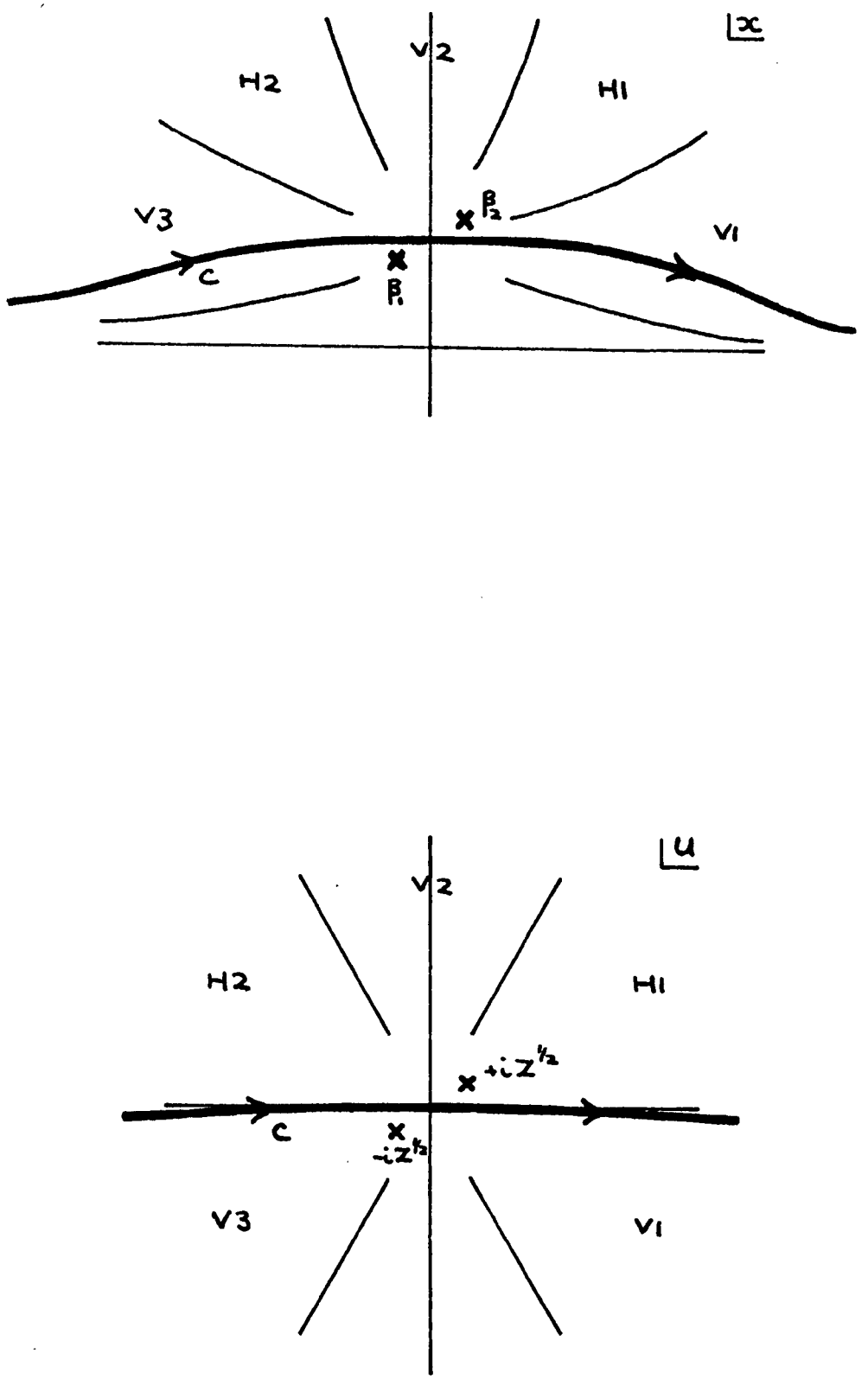


Figure 6.8 The Uniform Mapping from x to u near the Double Saddle

Substituting $u = \pm iZ^{\frac{1}{2}}$ in (6.10) gives

$$p_o = \{R_2 + R_1\}/2$$

$$q_o = \{R_2 - R_1\}/2iZ^{\frac{1}{2}}$$

As for isolated saddles, we shall assume that higher terms in (6.10) are negligible. Then after the transformation $u \to -u$

$$S_{\beta 12} \simeq \frac{1}{\sqrt{2\pi}} e^{iA(\underline{W})} \int_{-\infty}^{\infty} e^{i(u^3/3 + Z(\underline{W})u)} (p_o - uq_o)\, du$$

$$= \sqrt{2\pi}\, e^{iA(\underline{W})} \{p_o Ai(Z(\underline{W})) + iq_o Ai'(Z(\underline{W}))\} \qquad (6.11)$$

Actually at the double saddle, the formulae for p_o and q_o are indeterminate. Taking the limit, or re-evaluating them from their definition, gives

$$p_o = \left(\frac{-2}{\phi_3(\beta)}\right)^{1/3} = \frac{1}{(2W_3)^{1/3}}$$

$$q_o = \frac{-\phi_4(\beta)}{6\phi_3(\beta)} \left(\frac{-2}{\phi_3(\beta)}\right)^{2/3} = \frac{i}{(2W_3)^{7/6}}$$

As the saddles separate and Z increases, $S_{\beta 12}$ should go smoothly into the previous isolated saddle result. Z large and positive should give us only the contribution of β_1 (see figs.6.8 & 6.6a). The asymptotic forms of the Airy functions are

$$Ai(Z) \sim \frac{1}{2\sqrt{\pi}} Z^{-1/4} e^{-\frac{2}{3}Z^{3/2}}$$

$$Ai'(Z) \sim \frac{-1}{2\sqrt{\pi}} Z^{+1/4} e^{-\frac{2}{3}Z^{3/2}}$$

Substituting these into (6.11) gives exactly

$$\sqrt{\frac{i}{\phi_2(\beta_1)}}\, e^{i\phi(\beta_1)}$$

as required. Z large and negative should give us contributions from both β_1 and β_2 (see figs.6.8 & 6.6b), and substituting

$$Ai(Z) \sim \frac{1}{2\sqrt{\pi}} Z^{-1/4} \{e^{-\frac{2}{3}Z^{3/2}} + ie^{+\frac{2}{3}Z^{3/2}}\}$$

$$Ai'(Z) \sim \frac{-1}{2\sqrt{\pi}} Z^{+1/4} \{e^{-\frac{2}{3}Z^{3/2}} - ie^{+\frac{2}{3}Z^{3/2}}\}$$

into (6.10) gives exactly

$$\sqrt{\frac{i}{\phi_2(\beta_1)}}\, e^{i\phi(\beta_1)} + \sqrt{\frac{i}{\phi_2(\beta_2)}}\, e^{i\phi(\beta_2)}$$

again as required. The fact that (6.11) matches smoothly onto the isolated saddle results is probably the best justification for neglecting the higher order terms of (6.10). We only use the Airy approximation when the double saddle is far from the caustic, i.e. for fairly large $W_3 > 0$. The result is probably an asymptotic expansion in W_3 (c.f. values of p_0 and q_0 at the double saddle). Our approximation using only the first term of (6.10) seems to give good results where we use it, and this is probably because W_3 is large.

Very close to the double saddle we can make a transitional approximation by expanding the phase in a Taylor series to third order instead of using the mapping (6.6). We will deduce the transitional approximation from our uniform approximation, because it is much easier to handle analytically than the uniform approximation, and its derivation will shed some light on the behaviour of $Z(\underline{W})$, which will assist us to take the correct cube root.

Let us expand $\phi(x)$ about $\tilde{\beta} = (\beta_1 + \beta_2)/2$ up to cubic terms, so that it is antisymmetric about $\tilde{\beta}$, giving

$$\phi(x) \simeq \phi(\tilde{\beta}) + (x-\tilde{\beta})\phi_1(\tilde{\beta}) + (x-\tilde{\beta})^3\phi_3(\tilde{\beta})/6 \qquad (6.12)$$

Then $\phi_1(x) \simeq \phi_1(\tilde{\beta}) + (x-\tilde{\beta})^2\phi_3(\tilde{\beta})/2 = 0$ at $x = \beta_1$ and β_2

$$\Rightarrow \quad \phi_1(\tilde{\beta}) = -\frac{1}{2}\left(\frac{\beta_2-\beta_1}{2}\right)^2\phi_3(\tilde{\beta}) \qquad (6.13)$$

Putting (6.12) and (6.13) into (6.9) gives

$$Z^{\frac{1}{2}}(\underline{W}) \simeq \{-i\phi_3(\tilde{\beta})(\beta_2-\beta_1)^3/16\}^{1/3}$$

Now $\tilde{\beta}$ will be very close to the double saddle β, so we can take

$$\phi_3(\tilde{\beta}) \simeq \phi_3(\beta) = -4W_3$$

$$A(\underline{W}) = \tfrac{1}{2}\{\phi(\beta_1) + \phi(\beta_2)\} \simeq \phi(\beta,\underline{W})$$

$$p_0 \simeq (-2/\phi_3(\beta))^{1/3} = 1/(2W_3)^{1/3}$$

and if W_3 is sufficiently large q_0 is negligible.

Then

$$Z^{\frac{1}{2}}(\underline{W}) \simeq (iW_3/4)^{1/3}(\beta_2-\beta_1) = e^{i(\pi/6+2n\pi/3)}(W_3/4)^{1/3}(\beta_2-\beta_1)$$

and the u saddles are at

$$\pm iZ^{\frac{1}{2}}(\underline{W}) \simeq \pm e^{i2(1+n)\pi/3}(W_3/4)^{1/3}(\beta_2-\beta_1)$$

By choosing $n = -1$ we can put the u saddles at exactly the same angles relative to the contour and the double saddle as the x saddles. Then

clearly the way to choose the cube root in (6.9) is such that

$$\arg(\beta_2-\beta_1) - \pi/3 < \arg(iZ^{\frac{1}{2}}) < \arg(\beta_2-\beta_1) + \pi/3.$$

Since we shall only use the Airy approximation close to the double saddle, $(\beta_2-\beta_1)$ will not go significantly outside the first quadrant (see fig.6.5), and therefore we just require that

$$-\pi/12 < \arg(iZ^{\frac{1}{2}}) < \pi/2 + \pi/12, \quad \text{i.e.} \quad -\pi/2 - \pi/12 < \arg Z^{\frac{1}{2}} < \pi/12.$$

Our transitional approximation is

$$S \simeq \frac{\sqrt{2\pi}\, e^{i\phi(\beta,\underline{W})}}{(2W_3)^{1/3}} \operatorname{Ai}(-(W_3/4)^{2/3}(\beta_2(\underline{W})-\beta_1(\underline{W}))^2)$$

Let us use this to discuss the behaviour of S in the plane $W_2 = 0$ near the complex whisker, to discover whether there is any analogy to a real caustic here. Put $W_1 = (W_3/2)^2 + \xi$. Then

$$\beta_{1/2} = i\sqrt{\frac{W_3}{2}} \; -/+ \; i\sqrt{\frac{-\xi}{2W_3}} \quad \text{for} \quad \xi < 0$$

$$i\sqrt{\frac{W_3}{2}} \; -/+ \; \sqrt{\frac{\xi}{2W_3}} \quad \text{for} \quad \xi > 0 \qquad \text{(see fig.6.5a)}$$

$\tilde{\beta} = \beta = i\sqrt{\frac{W_3}{2}}$ giving

$$\phi(\beta,\underline{W}) = i\left\{\frac{8}{15}\left(\frac{W_3}{2}\right)^{5/2} + \xi\left(\frac{W_3}{2}\right)^{1/2}\right\}$$

and $\beta_2-\beta_1 = 2i\sqrt{-\xi/2W_3}$ or $2\sqrt{\xi/2W_3}$

$\Rightarrow$ $(\beta_2-\beta_1)^2 = 2\xi/W_3$ for all ξ giving

$$S \simeq \frac{\sqrt{2\pi}}{(2W_3)^{1/3}} e^{-\frac{8}{15}(W_3/2)^{5/2}} e^{-\xi(W_3/2)^{1/2}} \operatorname{Ai}(-\xi/(2W_3)^{1/3})$$

For $W_3 = 6$ this becomes

$$S \simeq 2.68\times10^{-4}\, e^{-1.73\xi} \operatorname{Ai}(-\xi/2.29).$$

In fig.6.9 we plot $\operatorname{Ai}(-\xi/2.29)$ to show where we might expect the caustic, and $e^{-1.73\xi}\operatorname{Ai}(-\xi/2.29)$ to show that the exponential decay completely swamps the Airy function leaving no trace of any amplitude maximum.

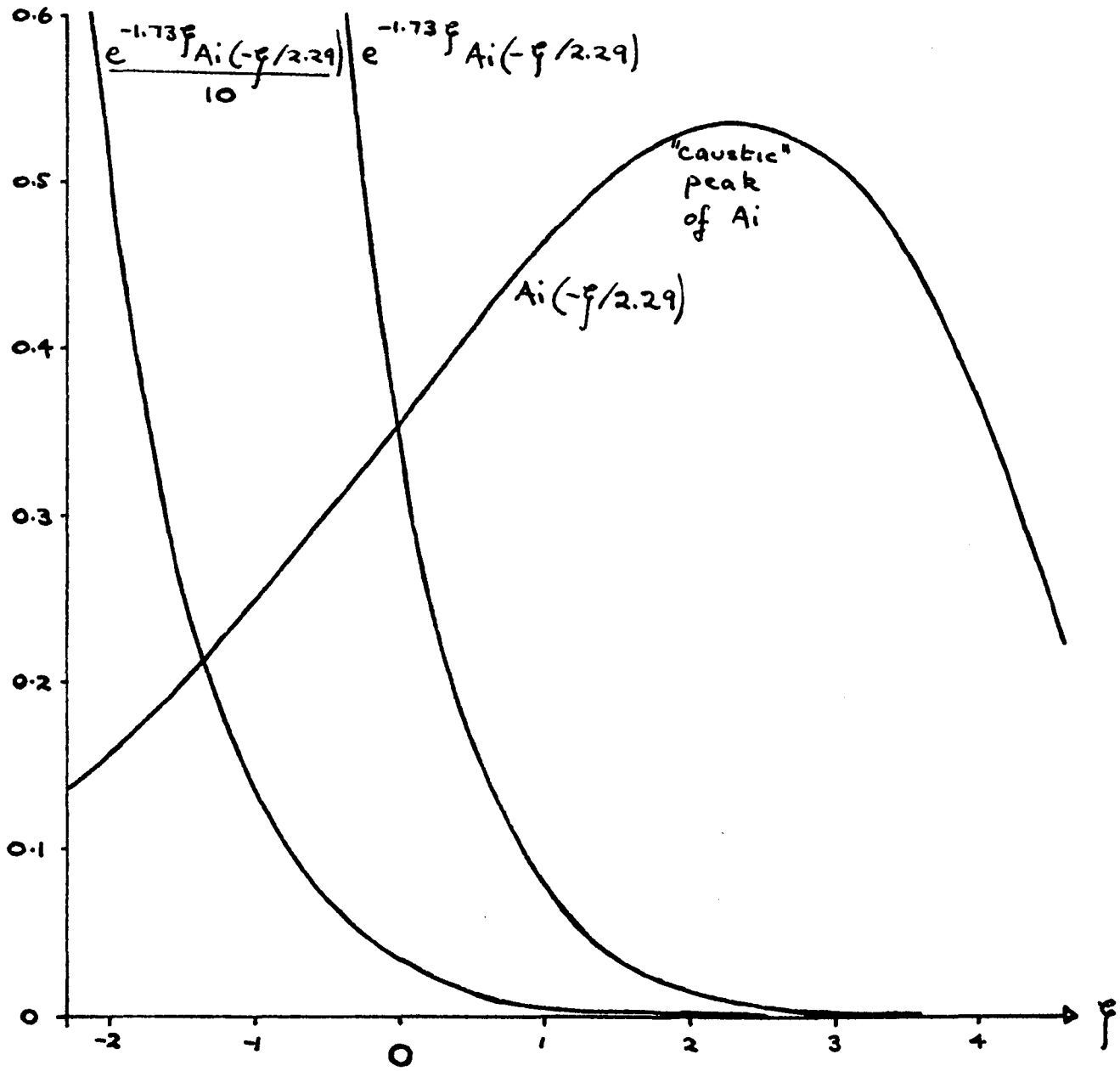


Figure 6.9 Amplitude through the Complex Whisker

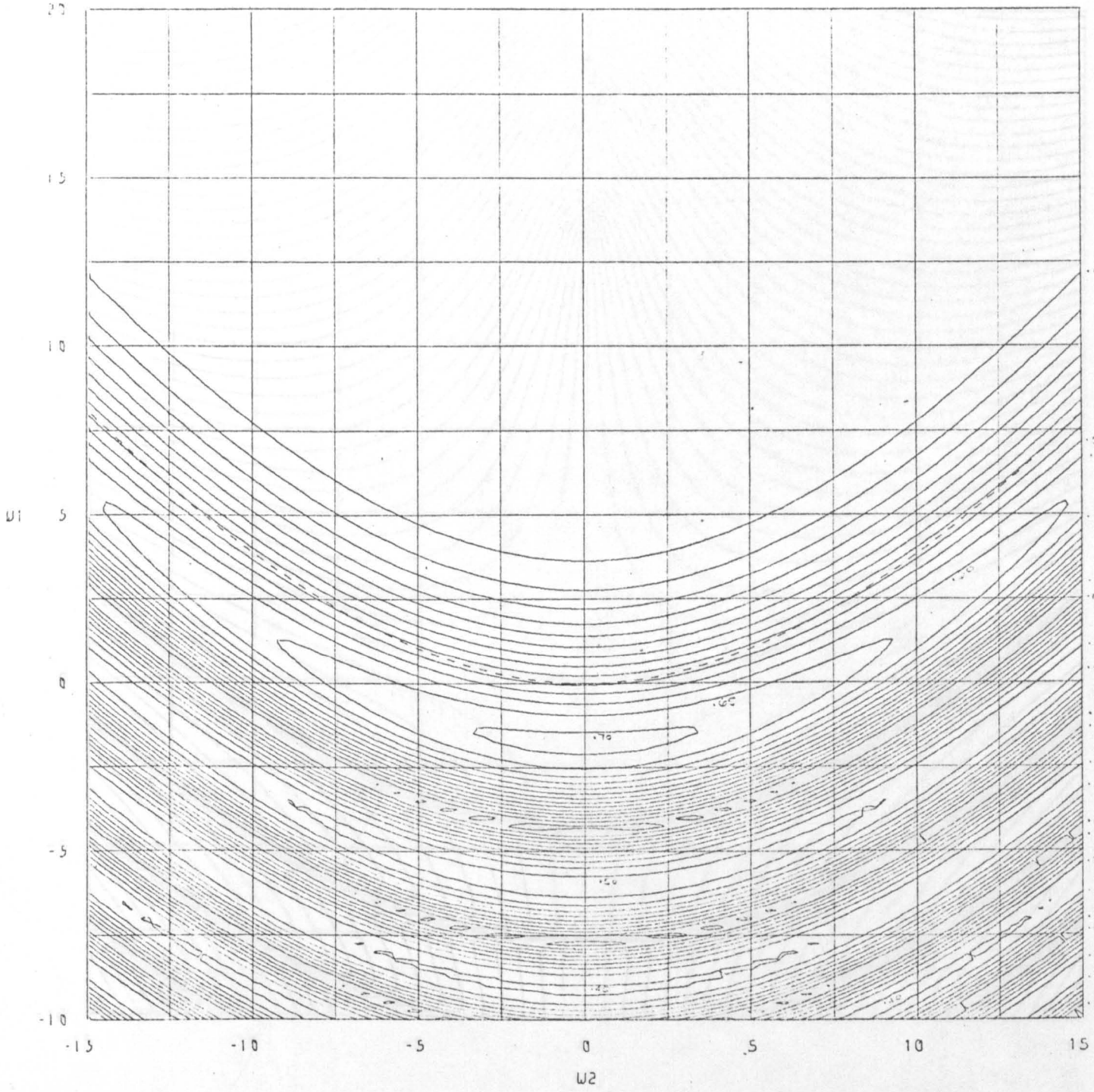
6.4 The Swallowtail Diffraction Pattern and its Nulls

We have now developed all the methods necessary to evaluate $S(W_1,W_2,W_3)$, and in figures 6.10 to 6.13 we display a few typical sections. The computational details are discussed in the appendix. No serious attempt has been made to label the contours because of their complexity. Probably the best way to do this would be to use shading techniques as in the next chapter. Also a more careful study is needed to resolve the detail for $W_3 < 0$, and the results presented really only constitute a preliminary survey. Experimental pictures of the swallowtail diffraction pattern are provided by Berry (1976), and it also arises in the interaction of the elliptic umbilic with a simple line caustic in the "triple junction" analysed by Berry & Nye (1977).

The pattern for $W_3 = +6$, shown in fig.6.10 is very simple. Essentially it is a series of light and dark fringes below and parallel to the caustic, whose amplitude is maximal on the mirror plane $W_2 = 0$, with a rapid decrease of amplitude above the caustic. This is just what one would expect from the shape of the caustic. The phase structure is more interesting, showing a sequence of nulls in the plane $W_2 = 0$ only. Notice that nulls only appear below the caustic and above the complex whisker, at $W_1 = 9$, as predicted in section 6.2. Four different integration algorithms are used in different regions above the caustic, and the smoothness of the phase lines shows how well they match up! The phase lines below the caustic are tending to coalesce along the dark fringes, as they would do completely in the Airy limit. This causes a noticeable difference in the phase structure around the nulls in the light and dark regions, which we shall investigate in detail later.

The pattern for $W_3 = 0$, shown if fig.6.11, is similar to that for $W_3 = +6$, but the fringes are closer together and are tending to have maximum amplitude away from $W_2 = 0$, in anticipation of the appearance of the cusps of the swallow's tail. (The apparent break-up of the fringes along their length is almost certainly a computational artefact). In fig.6.11b we plot $\log_e$(amp) to show the amplitude in the dark region. Since this is very similar for all W_3, we only display it in this one case. The complex whisker has now merged with the caustic, thereby allowing nulls in the dark region very close to the caustic.

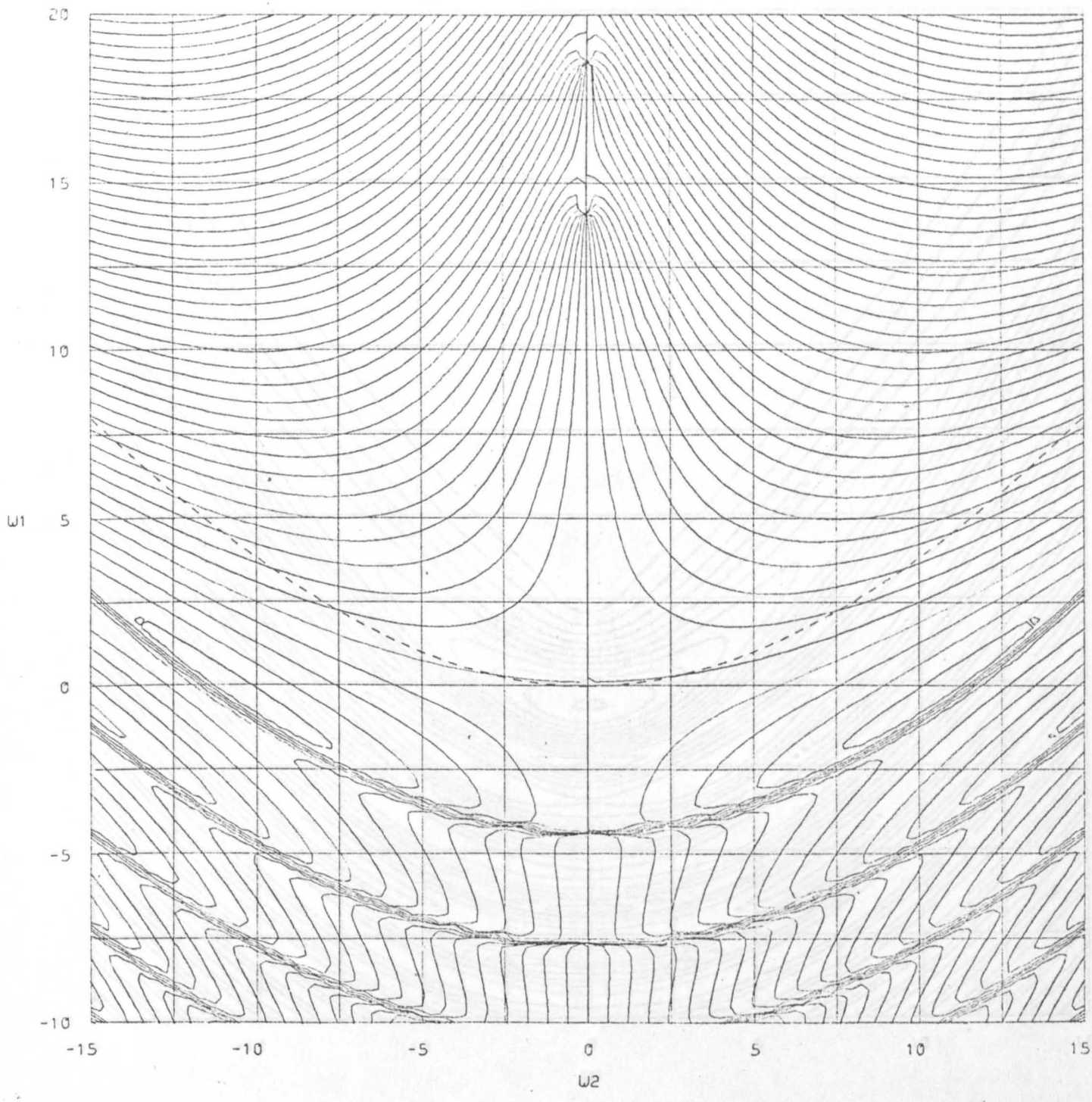
The pattern for $W_3 = -4$, shown in fig.6.12, has become so complicated that the contour intervals have been doubled. Well away from the caustic we still have simple fringes with maximum amplitude in



SWALLOWTAIL [W3=+6 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·05

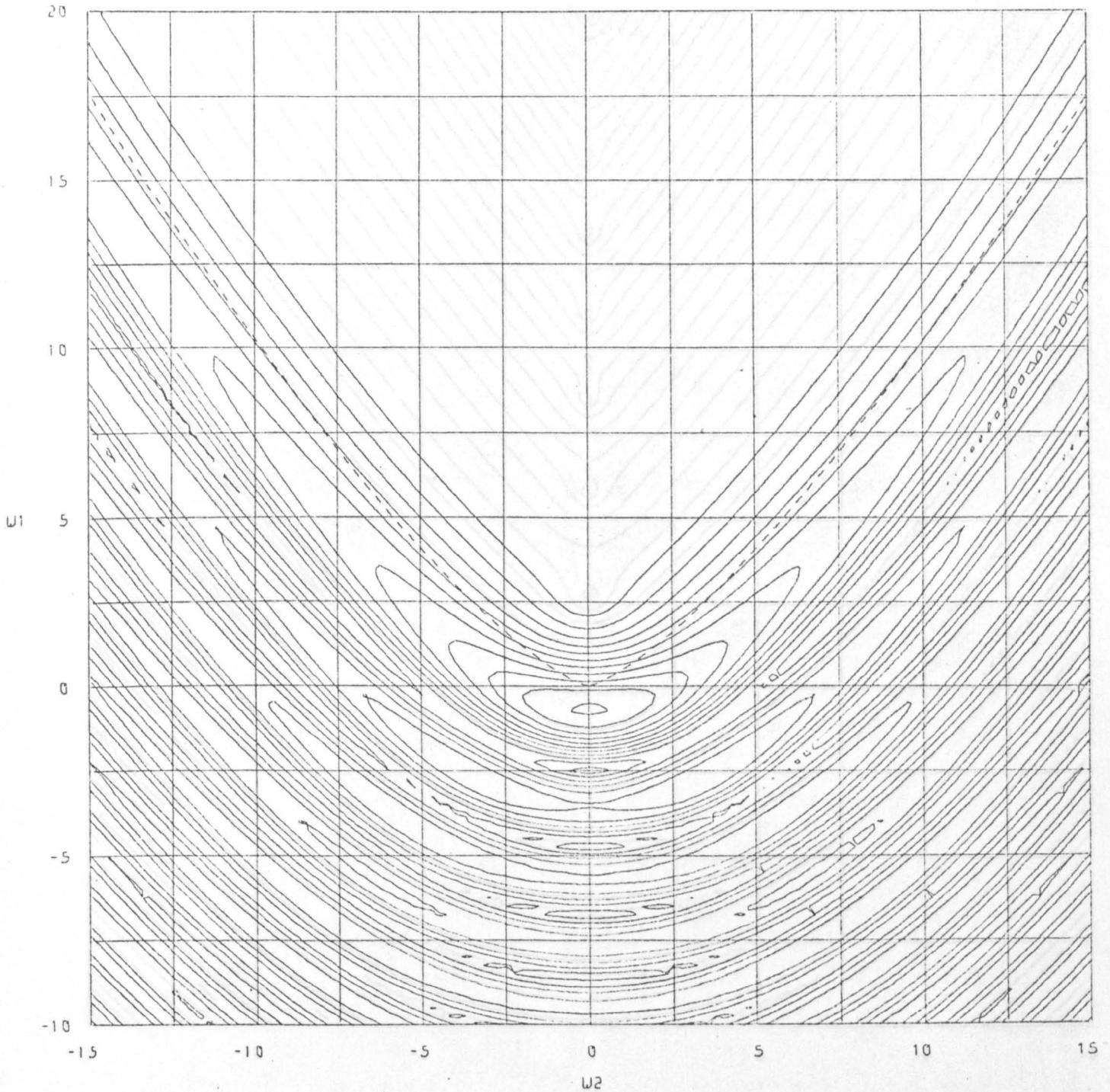
Figure 6.10a

Maximum Amplitude = 0.65



SWALLOWTAIL [W3=+6 SECTION]
CONTOURS OF PHASE in steps of Pi/8

Figure 6.10b



SWALLOWTAIL [W3=0 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·1

Figure 6.11a

Maximum Amplitude = 1.11

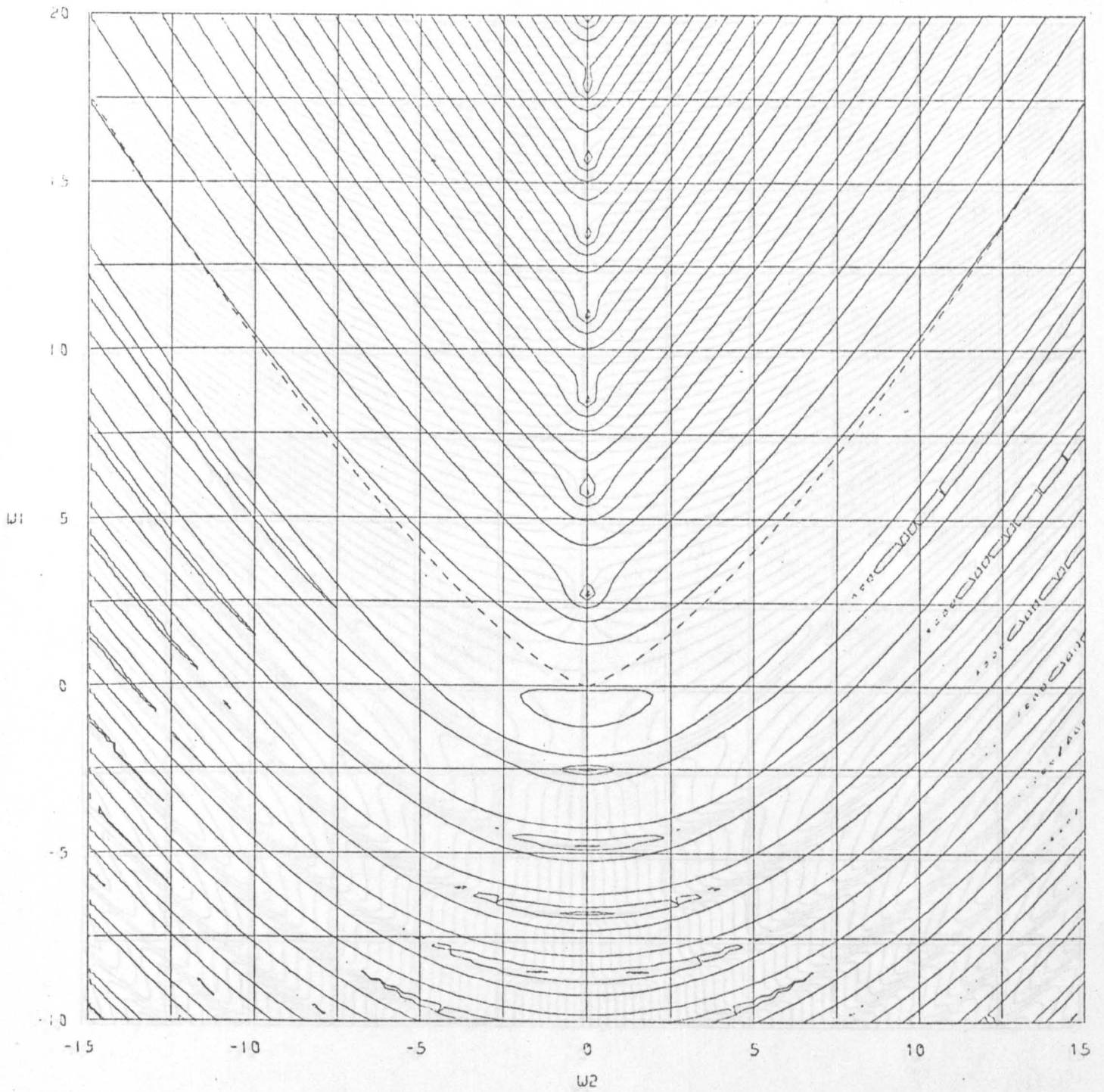
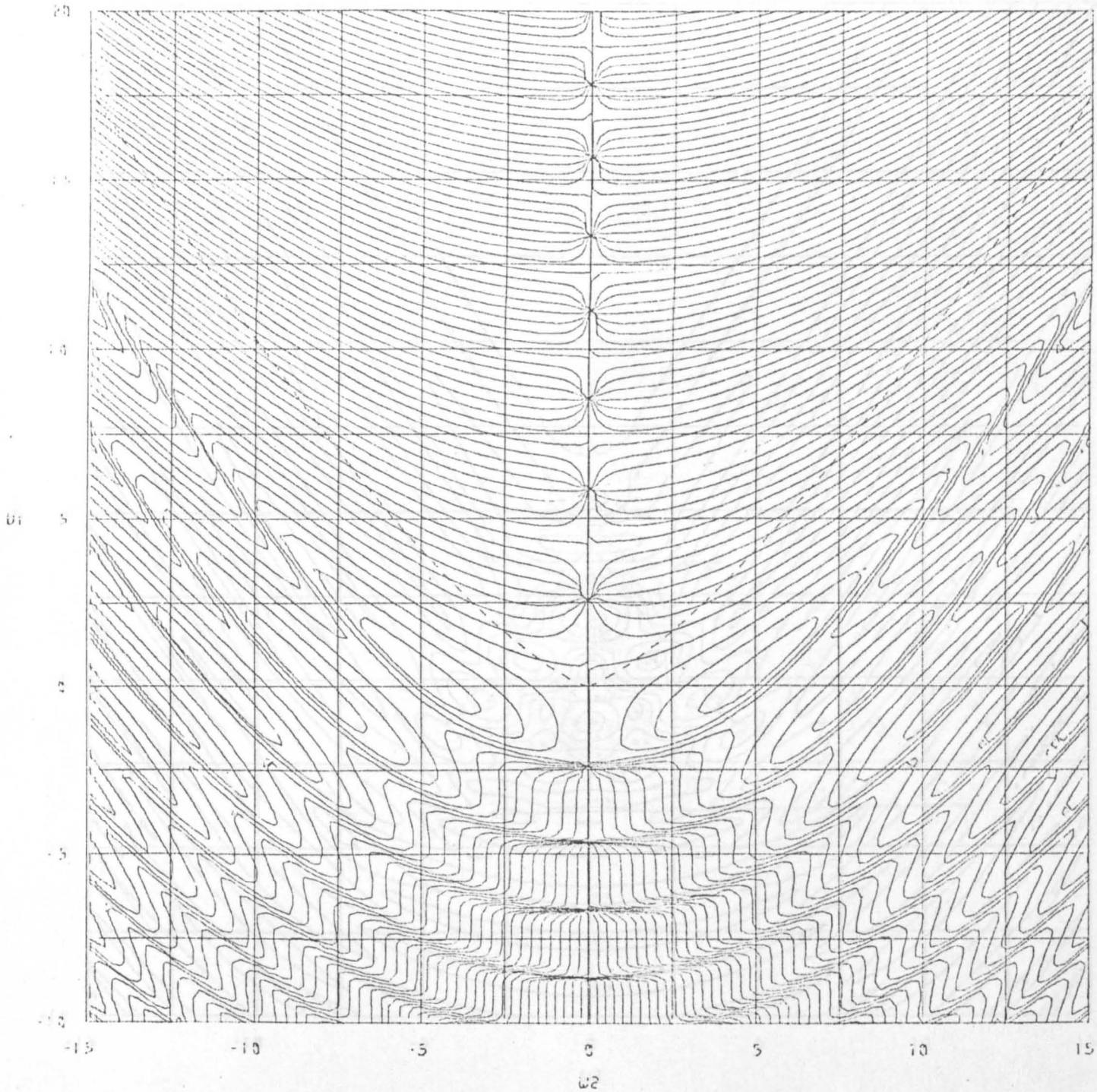
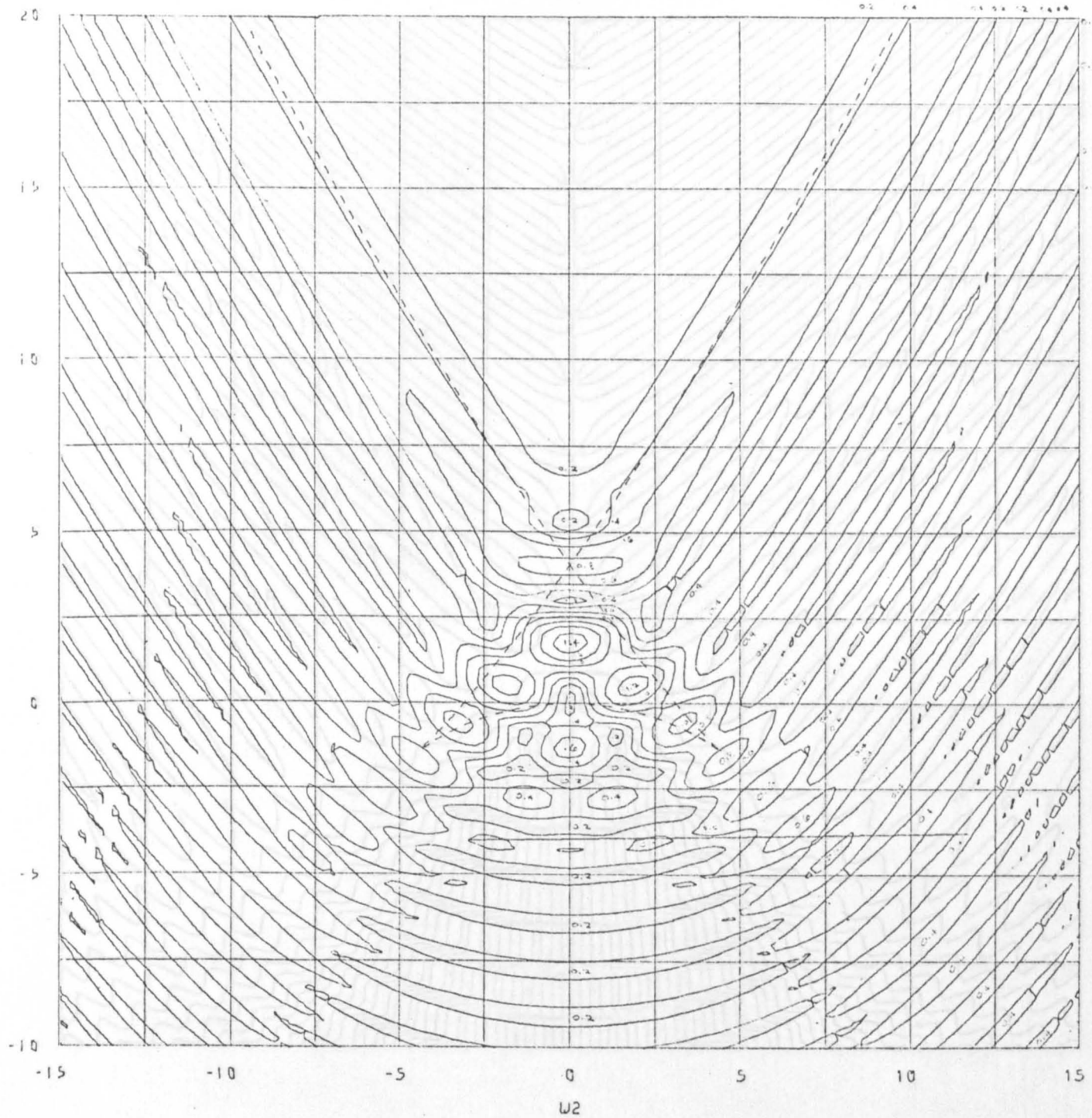

20
15
10
5
0
-5
-10
W1
-15
-10
-5
0
5
10
15
W2
SWALLOWTAIL [W3=0 SECTION]
CONTOURS OF LOGe[AMP] in steps of 1·0

Figure 6.11b



SWALLOWTAIL [W3=0 SECTION]
CONTOURS OF PHASE in steps of Pi/8

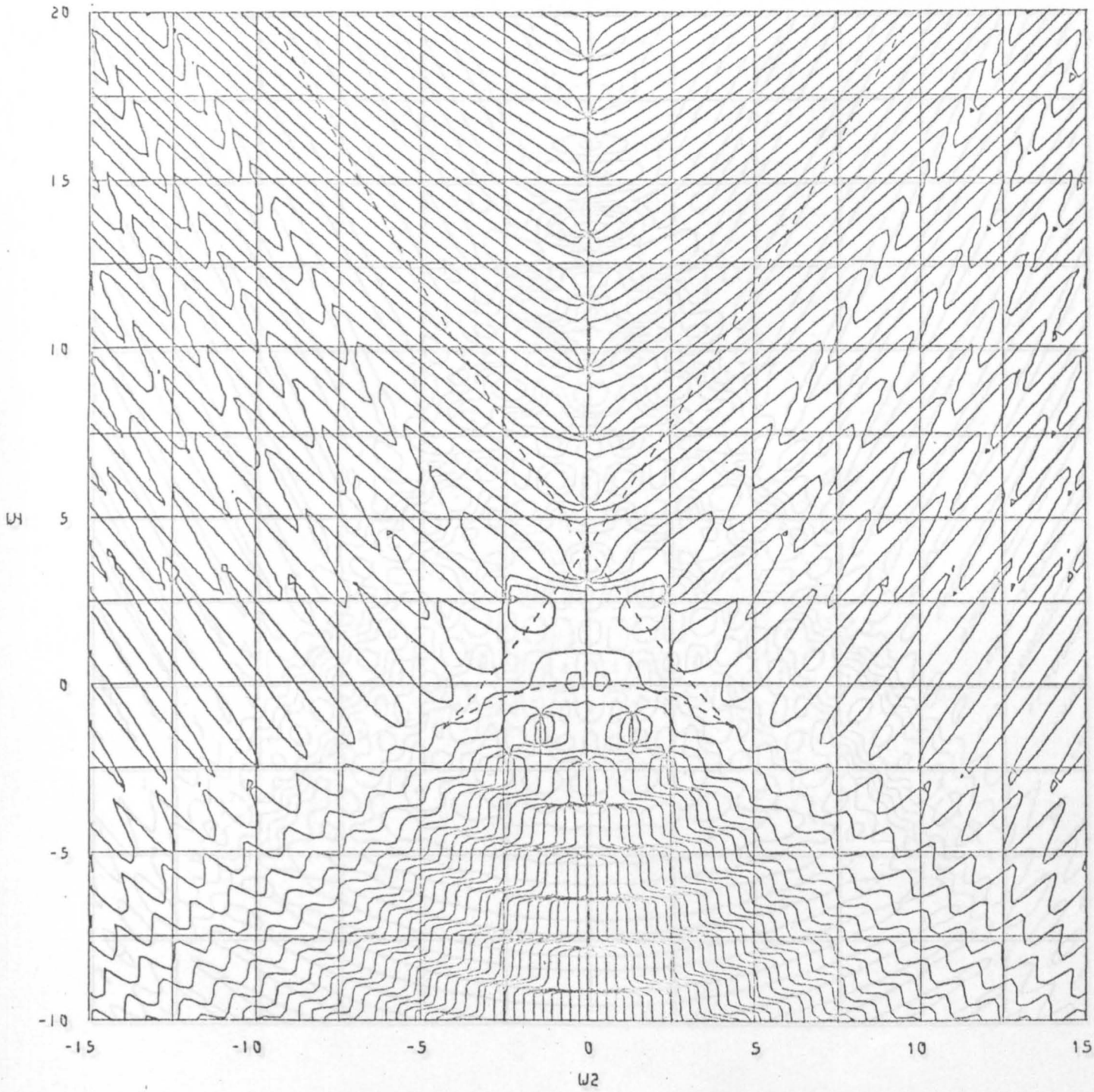
Figure 6.11c



SWALLOWTAIL [W3=-4 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·2

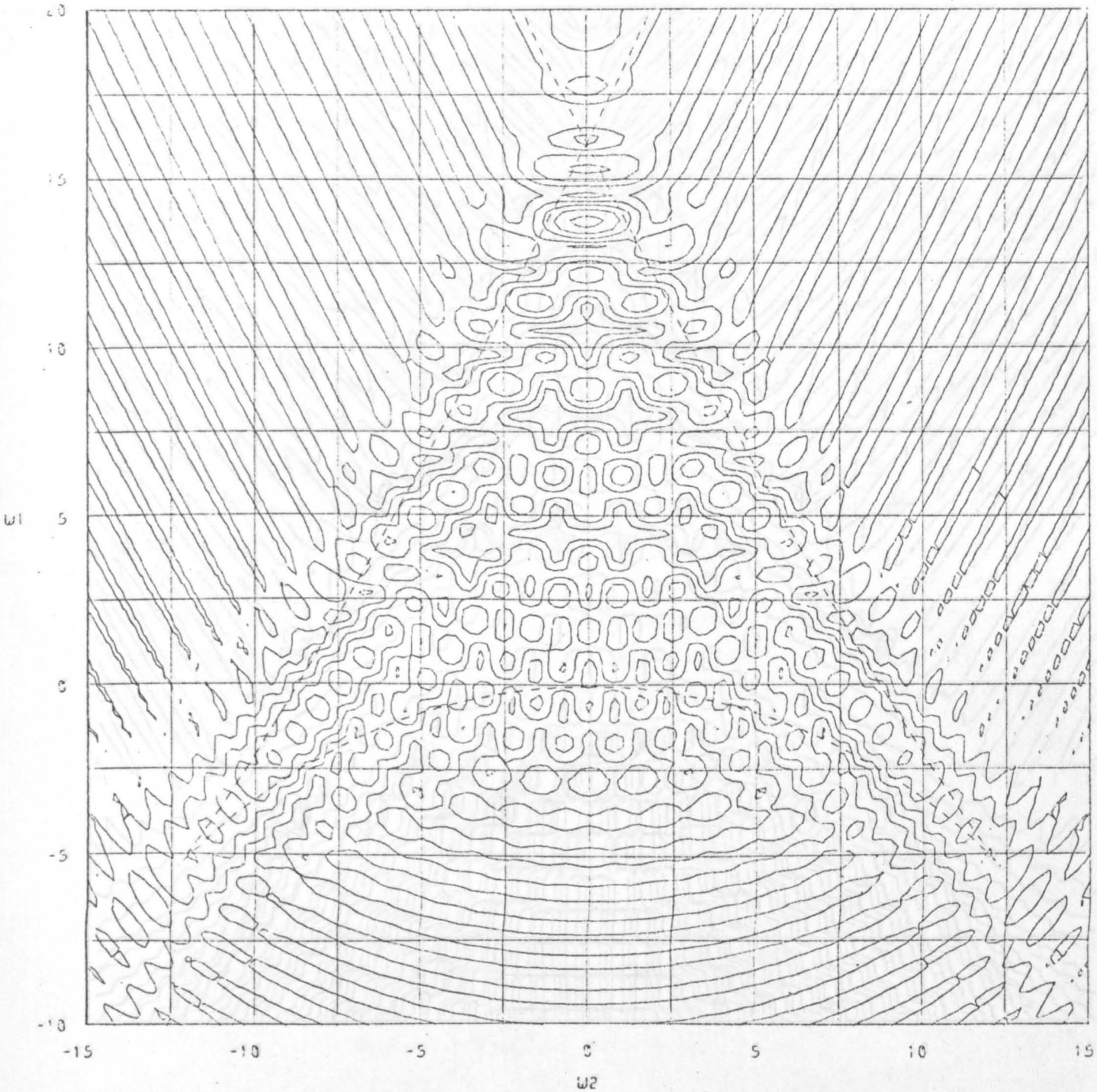
Figure 6.12a

Maximum Amplitude = 1.50



SWALLOWTAIL [W3=-4 SECTION]
CONTOURS OF PHASE in steps of PI/4

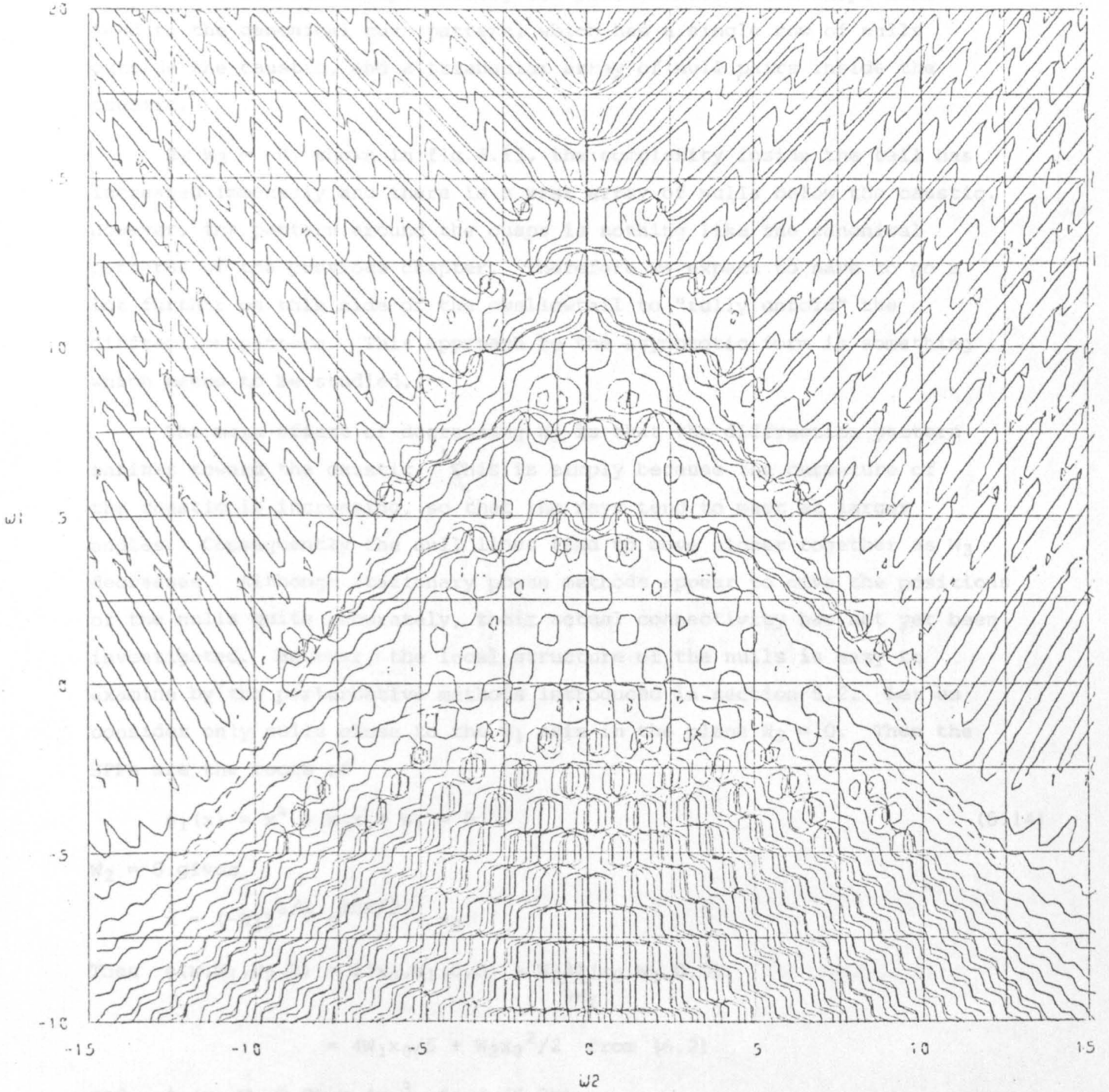
Figure 6.12b



SWALLOWTAIL [W3=-8 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·2

Figure 6.13a

Maximum Amplitude = 1.14



SWALLOWTAIL [W3=-8 SECTION]
CONTOURS OF PHASE in steps of Pi/4

Figure 6.13b

the direction of the cusps. Around the triangular tail the pattern has broken up into an array of spots, with the main peak of amplitude roughly in the middle. We now have nulls outside $W_2 = 0$. We know that these must occur, because asymptotically the pattern around the cusps must tend to the canonical cusp pattern, which has a single row of nulls outside the caustic, and a triangular array of null pairs inside the caustic.

By $W_3 = -8$, shown in fig.6.13, the complexity inside the tail has increased immensely and there is a vast array of nulls below the caustic. However, the pattern around the cusps is nothing like the canonical patterns of the previous chapter. Therefore we expect to have to go a lot further on this side of the swallowtail to "fully unfold" the diffraction pattern. This approach to the asymptotic form is something which needs to be studied.

The main effect of decreasing W_3 is that the diffraction pattern shrinks toward the caustic. This is simply because the curvature of the caustic is increasing, so that the rays tend to meet at larger angles. Consequently the null lines tend to move closer together as W_3 decreases. Although stationary phase methods appear to give the positions of the nulls quite accurately, their actual connectivity has not yet been investigated. However, the local structure of the nulls is easy to examine by the perturbative methods introduced in section 5.2. Let us consider only nulls close to the W_1 axis in the plane $W_3 = 0$. Then the SPPs are the roots of

$$\phi_1(x) = x^4 + W_2 x + W_1 = 0 \qquad (6.14)$$

$W_2 = 0$ gives

$$x = (-W_1)^{1/4}\, e^{in\pi/2} = x_0$$

Then $\phi(x,W_1,W_2,0) \simeq \phi(x_0,W_1,0,0) + W_2 \frac{\partial\phi}{\partial W_2}(x_0,W_1,0,0)$

$$= 4W_1x_0/5 + W_2x_0^2/2 \quad \text{from (6.2)}$$

and $\phi_2(x_0,W_1,0,0) = 4x_0^3$ from (6.3b)

For $W_1 < 0$ we have $x_0 = \pm(-W_1)^{1/4}, \pm i(-W_1)^{1/4}$. Ignoring the two complex saddles, we have

$$\phi \simeq \mp\frac{4}{5}(-W_1)^{5/4} + \frac{W_2}{2}(-W_1)^{1/2}$$

and $\phi_2 \simeq \pm 4(-W_1)^{3/4}$

giving

$$S \simeq \sqrt{\frac{1}{4(-W_1)^{3/4}}}\left\{e^{i\{-\frac{4}{5}(-W_1)^{5/4}+W_2(-W_1)^{1/2}/2+\pi/4\}}+e^{i\{+\frac{4}{5}(-W_1)^{5/4}+W_2(-W_1)^{1/2}/2-\pi/4\}}\right\}$$

$$= \frac{1}{(-W_1)^{3/8}}\, e^{iW_2(-W_1)^{1/2}/2}\cos\{\pi/4 - 4(-W_1)^{5/4}/5\}$$

and $|S|$ appears to be independent of W_2 to first order. This is one of the cases when the variation of the amplitude of the rays is important, since both rays are "in phase" with respect to W_2 variation. More accurately

$$\phi_2(x,W_1,W_2,0) = \phi_2(x_0,W_1,0,0) + \delta x(W_2)\phi_3(x_0,W_1,0,0) + W_2\frac{\partial\phi_2}{\partial W_2}(x_0,W_1,0,0)$$

$$+ O(W_2^2)$$

$$= 4x_0^3 + \delta x(W_2)12x_0^2 + W_2 + O(W_2^2)$$

For small W_2, let us expand x about x_0 as $x = x_0 + W_2a + O(W_2^2)$. If we avoid the neighbourhood of the caustic where $x_0 \simeq 0$, then equating coefficients in (6.14) gives

$$x = x_0 - W_2/4x_0^2$$

so that

$$\phi_2(x,W_1,W_2,0) = 4x_0^3 - 2W_2 + O(W_2^2)$$

and $$\phi_2^{-1/2} = \sqrt{\frac{1}{4x_0^3}}\left(1 + \frac{W_2}{4x_0^3}\right) + O(W_2^2)$$

For $x_0 = \pm(-W_1)^{1/4}$,

$$\phi_2^{-1/2} \simeq \sqrt{\frac{\pm 1}{4(-W_1)^{3/4}}}\left(1 \pm \frac{W_2}{4(-W_1)^{3/4}}\right)$$

giving

$$S \simeq \frac{e^{iW_2(-W_1)^{1/2}/2}}{(-W_1)^{3/8}}\left\{\cos\{\pi/4-4(-W_1)^{5/4}/5\} + \frac{W_2 i}{4(-W_1)^{3/4}}\sin\{\pi/4-4(-W_1)^{5/4}/5\}\right\} \quad (6.15)$$

For $W_1 > 0$ we have $x_0 = \pm e^{i\pi/4}W_1^{1/4}$, $\pm e^{-i\pi/4}W_1^{1/4}$. We know from section 6.2 that the contributing saddles are

$$\beta_{10} = e^{i3\pi/4}W_1^{1/4} \qquad \text{and} \qquad \beta_{20} = e^{i\pi/4}W_1^{1/4}$$

Then for $W_2 > 0$

$$\phi(\beta_1,W_1,W_2,0) \simeq 4e^{i3\pi/4}W_1^{5/4}/5 + W_2e^{i3\pi/2}W_1^{1/2}/2$$

$$\phi(\beta_2,W_1,W_2,0) \simeq 4e^{i\pi/4}W_1^{5/4}/5 + W_2e^{i\pi/2}W_1^{1/2}/2$$

Since now W_2 affects the two rays differently it should be sufficient to use the value of ϕ_2 at β_{10} and β_{20}, viz.

$$\sqrt{\frac{i}{\phi_2(\beta_{10})}} = \frac{e^{i\pi/8}}{2W_1^{3/8}} \quad \text{and} \quad \sqrt{\frac{i}{\phi_2(\beta_{20})}} = \frac{e^{-i\pi/8}}{2W_1^{3/8}}$$

where we have chosen the square roots to lie in the right hand half plane as prescribed in section 6.2. Then

$$S \simeq \frac{1}{2W_1^{3/8}} \left\{ e^{i\pi/8} e^{i\{(-1+i)\frac{4}{5\sqrt{2}} W_1^{5/4} - iW_2 W_1^{1/2}/2\}} + e^{-i\pi/8} e^{i\{(1+i)\frac{4}{5\sqrt{2}} W_1^{5/4} + iW_2 W_1^{1/2}/2\}} \right\}$$

$$= \frac{e^{-\frac{4}{5\sqrt{2}} W_1^{5/4}}}{W_1^{3/8}} \left\{ \cos(\pi/8 - \tfrac{4}{5\sqrt{2}} W_1^{5/4}) + i\frac{W_2}{2} W_1^{1/2} \sin(\pi/8 - \tfrac{4}{5\sqrt{2}} W_1^{5/4}) \right\} \qquad (6.16)$$

In the bright region nulls occur at

$$\cos(\pi/4 - 4(-W_1)^{5/4}/5) = 0 \qquad W_2 = 0$$

$$\Rightarrow \quad W_1 = -\{5\pi(n-1/4)/4\}^{4/5} \equiv W_{1_n}$$

Expand W_1 about W_{1_n} as $W_1 = W_{1_n} + y$. Then $\cos(\cdot) \simeq -(-1)^n (-W_{1_n})^{1/4} y$, and to first order in y and W_2, the term in braces in (6.15) is

$$(-1)^n(-W_{1_n})\{-y + iW_2/4(-W_{1_n})\} = i(-1)^n(-W_{1_n})^{1/4}\{W_2/4(-W_{1_n}) + iy\}$$

This is a conical null of the form x + iy stretched by $\beta = 4(-W_{1_n})$ along the W_2 axis, whose phase increases when encircled in a positive sense.

In the dark region nulls occur at

$$\cos(\pi/8 - 4W_1^{5/4}/5\sqrt{2}) = 0 \qquad W_2 = 0$$

$$\Rightarrow \quad W_1 = \{5\sqrt{2}\pi(n-3/8)/4\}^{4/5} \equiv W_{1_n}$$

Expanding about W_{1_n} to first order, the term in braces in (6.16) is

$$(-1)^n \frac{W_{1_n}^{1/4}}{\sqrt{2}} \left\{ y + i\frac{W_{1_n}^{1/4}}{\sqrt{2}} W_2 \right\} = i(-1)^n \frac{W_{1_n}^{1/4}}{\sqrt{2}} \left\{ \frac{W_{1_n}^{1/4}}{\sqrt{2}} W_2 - iy \right\}$$

This is a conical null stretched by $\beta = \sqrt{2}/W_{1_n}^{1/4}$ along the W_2 axis, whose phase decreases when encircled in a positive sense. We calculate a few values of W_{1_n} and β:

	Bright region		Dark region	
n	W_n	β	W_n	β
1	-2.37	9.49	2.71	1.10
2	-4.67	18.69	5.81	0.91
3	-6.71	26.84	8.53	0.83
4	-8.60	34.40	11.04	0.78

Comparison with fig.6.11 shows that the positions of the nulls are given very well. The dark region nulls are not much distorted from conical nulls. However, the bright region nulls are squashed by a large factor in the vertical direction, so that the phase saddle point becomes very close to the actual null. This explains the structures we see in fig.6.11. The distortion of both sets of nulls increases away from the bifurcation set. The shape of the nulls depends on the way the amplitudes of the two interfering rays vary with W_2. In the bright region the amplitudes vary very slowly with W_2 through the factor $1/\sqrt{\phi_2}$, and hence the null is elongated in the W_2 direction. But in the dark region the amplitude varies exponentially with W_2 through the factor $\exp(-\mathrm{Im}\phi)$. This is comparable with the variation with W_2 through $\exp(i\mathrm{Re}\phi)$, so that the nulls are not greatly distorted, and arises because the rays are complex.

APPENDIX

A6.1 Computational Details

The behaviour of the integral for $W_3 \geq 0$ and for $W_3 < 0$ is sufficiently different that it is best to treat the two cases separately. For $W_3 \geq 0$ there may be one or two complex rays above the caustic, and two real rays below, while for $W_3 < 0$ there are always two complex rays above the caustic, but there may be two or four real rays below. We shall consider $W_3 \geq 0$ first, and examine the behaviour of

$$\phi_1(x,W_1,W_2,W_3) = x^4 + W_3x^2 + W_2x + W_1$$

for $W_2 \geq 0$, which is sketched in fig.A6.1.1.

As we decrease W_1 a pair of real SPPs appear and separate. Starting in the dark region, we must use the various steepest descent formulae until the amplitude has increased to a reasonable value (of 0.01). We start with $W_2 = 0$, $W_1 = \max.$ (and assume we are above the complex whisker, i.e. $W_1 > (W_3/2)^2$) and compute β_1 and β_2 from the simple quadratic formula. We then successively decrease W_1 and compute the new β_1 and β_2 by Newton's iteration from the previous values. Once we have crossed SB we only compute β_1. At the top of each W_1 scan we save the values of β_1 and β_2 to use as initial values for the next W_1 scan. At each point we check the convergence condition (6.5) for both saddles, and if either fails we use the uniform Airy approximation, otherwise we use the isolated saddle formulae. We check whether we have crossed SB, and if so we use only β_1, otherwise we use β_1 and β_2. On leaving the Airy region we know that we have crossed SB, and we also check whether we have crossed SB horizontally at the start of a W_1 scan. The integration methods used in different regions are summarised in fig.A6.1.2. At the double saddle (on the complex whisker) we must use the special Airy formula, and reset β_1 and β_2 afterwards to preserve their identity correctly. Ai(Z) and Ai'(Z) are computed from their convergent series expansions.

When the amplitude has reached 0.01 on each W_1 scan we switch to the normal quadrature plus asymptotic series method. We compute the position of the minimum of ϕ_1, i.e. the single real root of

$$\phi_2 = 4x^3 + 2W_3x + W_2 = 0$$

using the standard formulae for the roots of a cubic: call it x_{min}. This is where the pair of real SPPs will appear as we cross the caustic. We set our initial cutoffs A and B such that on the caustic

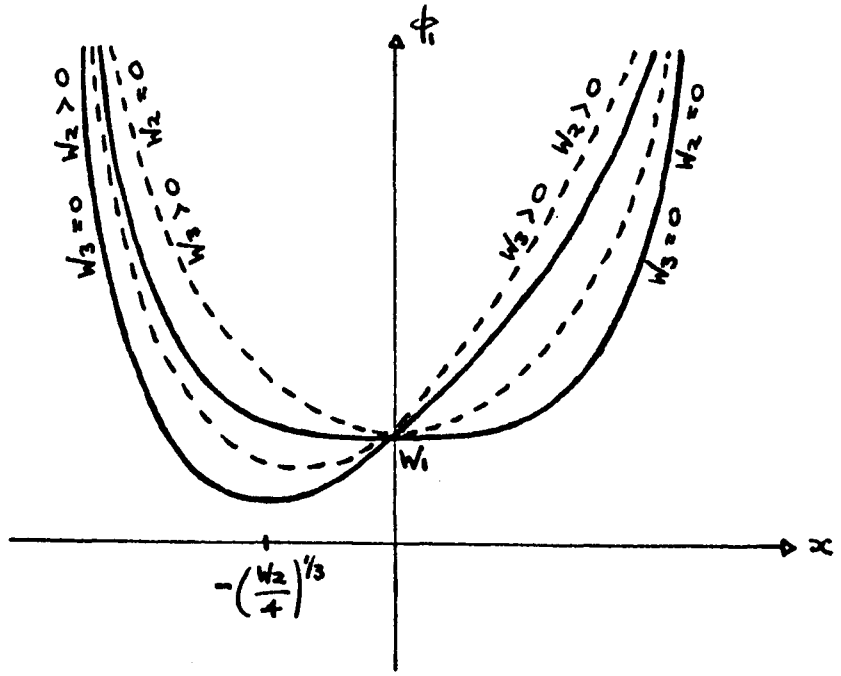


Figure A6.1.1 ϕ_1 for $W_3 > 0$

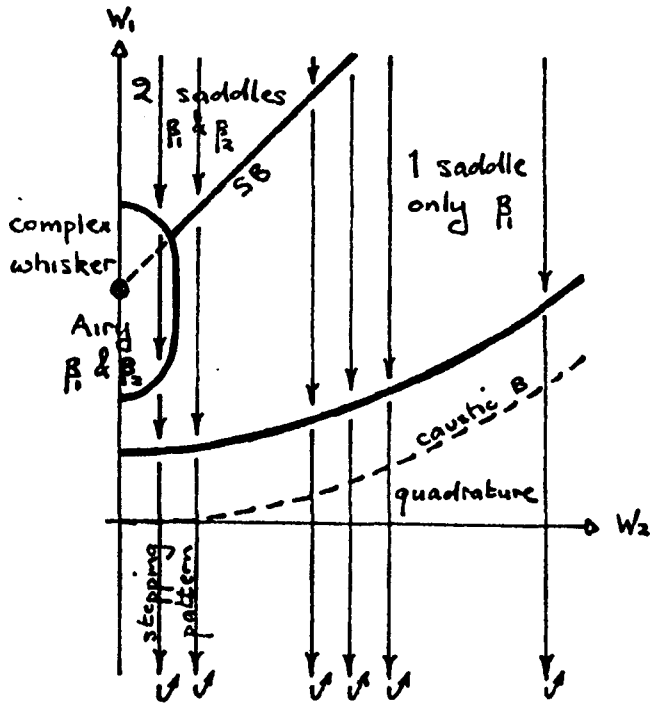


Figure A6.1.2 The Four Integration Regions

$\phi_1(A) \simeq \phi_1(B) \simeq C1$ by making a quadratic approximation to ϕ_1, giving

$$\begin{Bmatrix} A \\ B \end{Bmatrix} = x_{min} \pm \sqrt{\frac{C1}{\phi_3(x_{min})}}$$

We can compute the value of W_1 on the bifurcation set using (6.3a) (or (6.4a)), and then step the cutoffs out for convergence as usual (see appendix A5.1) in a dummy run. We then use these cutoffs in the dark region for continuity through the caustic, stepping out as necessary below the caustic.

For $W_3 < 0$, ϕ_1 is more complicated, as shown in fig.A6.1.3. In the dark region we use the two saddle formula until the amplitude $\geqslant 0.01$, as for $W_3 \geqslant 0$. We compute the positions of the minima of ϕ_1 as above. If there is only one we proceed as for $W_3 \geqslant 0$. If there are three we call their maximum and minimum values x_{max} and x_{min} respectively, and find the initial cutoffs around x_{min} as before. We define BIF as

$$BIF = x_{max}^4 + W_3 x_{max}^2 + W_2 x_{max}$$

which is the value of W_1 at which we encounter the second (higher) minimum of ϕ_1, i.e. cross the caustic for the second time. When $W_1 - BIF \leqslant C1$ we set $A = x_{max}$ and step out from there as necessary so as to be outside the new SPPs. For very small values of W_2, A will jump outside the second minimum immediately, and in particular when $W_2 = 0$ and ϕ_1 is symmetrical, the cutoffs will be forced to be symmetrical.

A Chebyschev integration routine was used. This was satisfactory for $W_3 \geqslant 0$, but for $W_3 < 0$ it was necessary to split up the range of integration in a fairly arbitrary manner to achieve convergence. It would be much better to use the quadrature routine discussed in appendix A7.1. For an error in each series of less than 0.005, the asymptotic forms in section 4.5 give the cutoffs as ±1.89, and the magnitudes of the first 3 terms of the series as

0.08, 0.013 and 0.005

In practice it was found satisfactory to use C1 = 5, and require the second and third terms to be less than 0.05 and 0.005 respectively. This low series accuracy was used to keep the cutoffs as close together as possible to try to avoid spurious convergence of the quadrature. With a better quadrature routine the series error could be decreased.

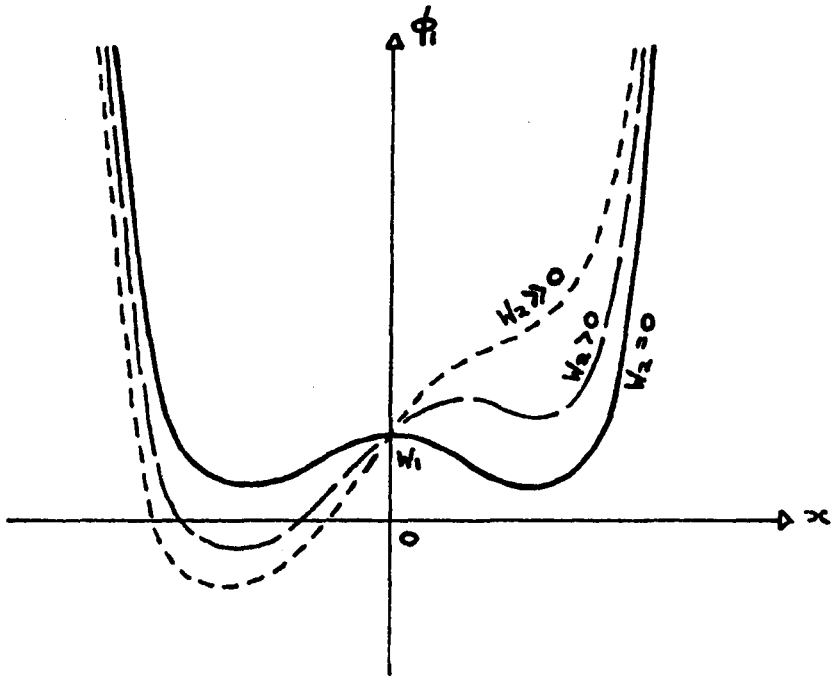


Figure A6.1.3 ϕ_1 for $W_3 < 0$

We plot the caustic as a dashed line using the parametric equations (6.4). The arc length s is given by

$$ds/dx = \pm 2(6x^2 + W_3)\sqrt{x^2 + 1}$$

therefore we find the coordinates of the ends of the dashes from

$$x_{n+1} = x_n - \delta s/2\,|6x^2 + W_3|\sqrt{x^2 + 1}, \qquad x_0 = 0$$

substituted into (6.4). For $W_3 > 0$ there is no problem. For $W_3 = 0$ the formula diverges at $x = 0$. But then $dW_2 = ds$

$$\Rightarrow \quad (W_2)_1 = -4x_1^3 = \delta s \quad \Rightarrow \quad x_1 = -(\delta s/4)^{1/3}$$

which we use to start the iteration. For $W_3 < 0$ the formula diverges at the cusps at $x^2 = -W_3/6$. When x_{n+1} becomes $< -\sqrt{-W_3/6}$ we set $x_{n+1} = -\sqrt{-W_3/6}$. At the cusp

$$ds/dW_2 = \pm\sqrt{1 - W_3/6}$$

so we take

$$\delta W_2 = +\delta s/\sqrt{1 - W_3/6}$$

after the cusp. But

$$\delta W_2 = -4(3x + \delta x)\delta x^2$$

(exactly) at the cusp, giving

$$\delta W_2 = +12\sqrt{-W_3/6}\ \delta x^2 + O(\delta x^3).$$

Equating these two expressions for δW_2 gives

$$\delta x \simeq -\sqrt{\frac{\delta s}{12\sqrt{(W_3/6-1)W_3/6}}}$$

which we use for the first step after the cusp. We plot (W_1, W_2) and $(W_1, -W_2)$ to give the whole caustic.

CHAPTER 7

THE ELLIPTIC UMBILIC DIFFRACTION PATTERN

7.1 The Canonical Integral and Caustic

In this chapter we study the double integral

$$E(W_1,W_2,W_3) = \frac{1}{2\pi}\iint_{-\infty}^{\infty} dx.\,dy\; e^{i\phi(x,y,W_1,W_2,W_3)} \tag{7.1}$$

where $\phi(x,y,W_1,W_2,W_3) = x^3 - 3xy^2 - W_3(x^2+y^2) - W_1x - W_2y$ (7.2)

We now have a two-dimensional state space, hence the double integral. It has a number of interesting analytical properties which we shall consider first. The transformation $y \to -y$ shows that

$$E(W_1,-W_2,W_3) = E(W_1,W_2,W_3)$$

and the transformation $x \to -x$, $y \to -y$ shows that

$$E(W_1,W_2,-W_3) = E^*(W_1,W_2,W_3).$$

Therefore $E(W_1,W_2,0)$ is pure real, which implies that generically zeros of E in the plane $W_3 = 0$ will occur as lines, whereas the zeros in any other plane will be points. Therefore we expect the null lines of E to be a set of twisted space curves plus a set of plane curves lying in the singular plane $W_3 = 0$.

If we rotate the W_1,W_2 axes through an angle θ, then by rotating the xy axes through the same angle we can regain the original integral if $\theta = \pm 2\pi/3$. Therefore E has a triad rotation axis along W_3, plus three mirror planes in the plane $W_2 = 0$ and its symmetry relatives. So we only need to compute E in a 60^0 sector with $W_3 \geqslant 0$. The caustic must display at least as high a symmetry as the wavefunction. It is given by the ray equations

$$\left.\begin{aligned} \partial\phi/\partial x &= 3x^2 - 3y^2 - 2W_3x - W_1 = 0 \\ \partial\phi/\partial y &= -6xy - 2W_3y - W_2 = 0 \end{aligned}\right\} \tag{7.3}$$

plus the coalescence condition

$$H\phi = \det\begin{vmatrix} \phi_{xx} & \phi_{xy} \\ \phi_{yx} & \phi_{yy} \end{vmatrix} = 0$$

$$\Rightarrow \quad x^2 + y^2 = (W_3/3)^2 \tag{7.4}$$

(c.f. eqns.(4.4)). We can conveniently satisfy (7.4) by parametrising

x and y as

$$x = \frac{-W_3}{3}\cos\theta\,, \qquad y = \frac{-W_3}{3}\sin\theta$$

Then (7.3) gives

$$\left.\begin{aligned} W_1 &= \frac{W_3^2}{3}(\cos 2\theta + 2\cos\theta) \\ W_2 &= \frac{W_3^2}{3}(-\sin 2\theta + 2\sin\theta) \end{aligned}\right\} \tag{7.5}$$

with which we can easily plot sections W_3 = constant through the caustic, as in fig.7.1a. Each of these sections is the well known hypocycloid of three cusps, which is the locus of a point on the circumference of a circle which rolls around the inside of a circle having three times its radius. θ is the angle through which the centre of the moving circle has moved relative to the centre of the fixed circle. Fig.7.1b shows a sketch of the whole caustic. There are four real rays inside the caustic and two outside. Therefore there is no dark region, and we may expect nulls to occur anywhere (unlike the swallowtail).

The elliptic umbilic integral is closely related to the hyperbolic umbilic integral (from chapter 4):

$$H(W_1,W_2,W_3) = \frac{1}{2\pi}\iint_{-\infty}^{\infty} dx\,dy\; e^{i(x^3 + y^3 + W_3xy - W_1x - W_2y)} \tag{7.6}$$

since if we rotate the xy axes through $-\pi/4$ and rescale them thus

$$\left.\begin{aligned} x &\to 2^{1/6}(x - y)/\sqrt{2} \\ y &\to 2^{1/6}(x + y)/\sqrt{2} \end{aligned}\right\} \tag{7.7}$$

we find

$$H(W_1,W_2,W_3) = \frac{2^{1/3}}{2\pi}\iint_{-\infty}^{\infty} dx\,dy\,\exp i\Big(x^3 + 3xy^2 + \frac{W_3}{2^{2/3}}(x^2-y^2) - \frac{(W_1+W_2)}{2^{1/3}}x - \frac{(-W_1+W_2)}{2^{1/3}}y\Big)$$

The only essential difference between the elliptic and hyperbolic integrals is the sign of the y^2 terms (as implied by their names!). Now by expanding the cross term in (7.6) as

$$e^{iW_3xy} = \sum_{n=0}^{\infty}\frac{(iW_3xy)^n}{n!}$$

we may write (7.6) as

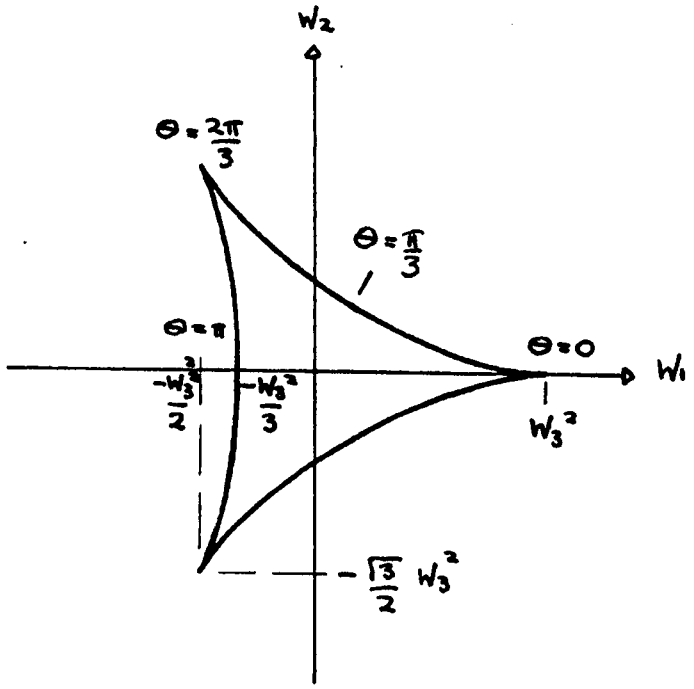


Figure 7.1a W_3 = constant section through the Caustic

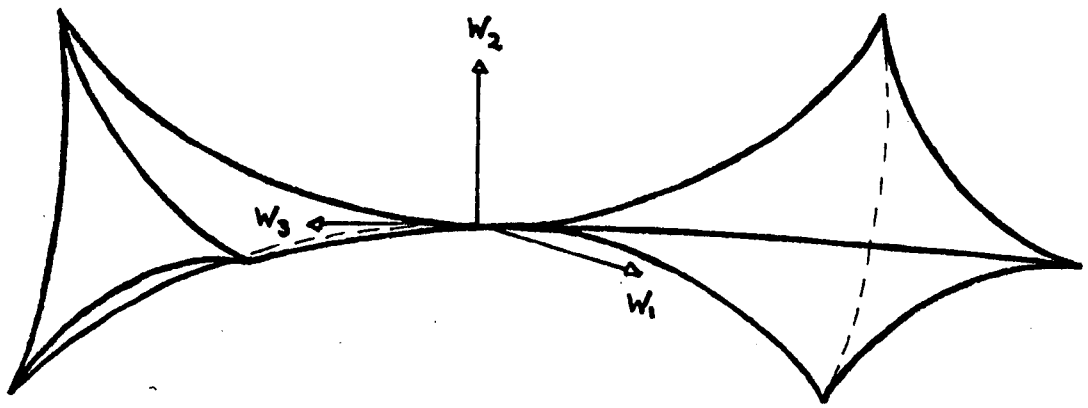


Figure 7.1b The Elliptic Umbilic Caustic

$$H(W_1,W_2,W_3) = 2\pi \sum_{n=0}^{\infty} \frac{(-iW_3)^n}{n!} \frac{1}{2\pi} \int_{-\infty}^{\infty} dx\,(ix)^n e^{i(x^3-W_1x)} \frac{1}{2\pi} \int_{-\infty}^{\infty} dy\,(iy)^n e^{i(y^3-W_2y)}$$

$$= \frac{2\pi}{3^{2/3}} \sum_{n=0}^{\infty} \frac{(-i3^{\frac{2}{3}}W_3)^n}{n!} \mathrm{Ai}^{(n)}\left(\frac{-W_1}{3^{\frac{1}{3}}}\right) \mathrm{Ai}^{(n)}\left(\frac{-W_2}{3^{\frac{1}{3}}}\right) \qquad (7.8)$$

where $\mathrm{Ai}(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, e^{i(t^3/3 + zt)}$ is the Airy integral function,

and $\mathrm{Ai}^{(n)} \equiv \frac{d^n \mathrm{Ai}(z)}{dz^n}$. Thus we have a convergent series expansion for H (see Connor (1973)), and the singular section $W_3 = 0$ takes the particularly simple form

$$H(W_1,W_2,0) = \frac{2\pi}{3^{2/3}} \mathrm{Ai}\left(\frac{-W_1}{3^{\frac{1}{3}}}\right) \mathrm{Ai}\left(\frac{-W_2}{3^{\frac{1}{3}}}\right).$$

Berry conjectured that a similar simple form should exist for the elliptic umbilic. To show that it does we must cast (7.1) into "pseudo-hyperbolic" form by changing the sign of y^2 relative to x^2. We wish to rotate the x and y contours by making the transformation

$$x \to \alpha x \qquad \text{and} \qquad y \to \beta y$$

such that $x^3 \to x^3$ and $xy^2 \to -xy^2$. Then either

$$\alpha = 1, \qquad \beta = \pm i \qquad (7.9a)$$

or $$\alpha = e^{\pm i2\pi/3}, \qquad \beta = e^{\pm i\pi/6} \qquad (7.9b)$$

Examination of $|e^{ix^3}|$ shows that asymptotically the x-contour Cx may only lie in the three sectors labelled (1), (2) & (3) in fig.7.2. Thus the real axis is equivalent to C_{21} or $C_{23} + C_{31}$. When x lies asymptotically anywhere in one of its allowed sectors, the sectors of the y plane shown hatched in fig.7.3 are forbidden to the y-contour Cy, because $|e^{-ixy^2}|$ would diverge. We cannot make the same rotation of Cy in sectors (1) & (2), i.e. along the whole of C_{21}, and therefore we cannot make the simpler transformation (7.9a). But we can use $\beta = e^{i\pi/6}$ everywhere along $C_{23} = -e^{i2\pi/3}C_{21}$, and $\beta = e^{-i\pi/6}$ everywhere along $C_{31} = -e^{-i2\pi/3}C_{21}$, which is precisely transformation (7.9b). Therefore we distort Cx into $C_{23} + C_{31}$, and put $y = e^{i\pi/6}Y$ on C_{23} and $y = e^{-i\pi/6}Y$ on C_{31}. Then using $C_{23} = -e^{i2\pi/3}C_{21}$ and $C_{31} = -e^{-i2\pi/3}C_{21}$ and replacing C_{21} by the real axis again gives

$$E(W_1,W_2,W_3) = -\frac{e^{i5\pi/6}}{2\pi} \iint_{-\infty}^{\infty} dX\, dY\, e^{i\{X^3+3XY^2-W_3e^{i\pi/3}(-X^2+Y^2)-W_1e^{i2\pi/3}X-W_2e^{i\pi/6}Y\}}$$

$$-\frac{e^{-i5\pi/6}}{2\pi} \iint_{-\infty}^{\infty} dX\, dY\, e^{i\{X^3+3XY^2-W_3e^{-i\pi/3}(-X^2+Y^2)-W_1e^{-i2\pi/3}X-W_2e^{-i\pi/6}Y\}}$$

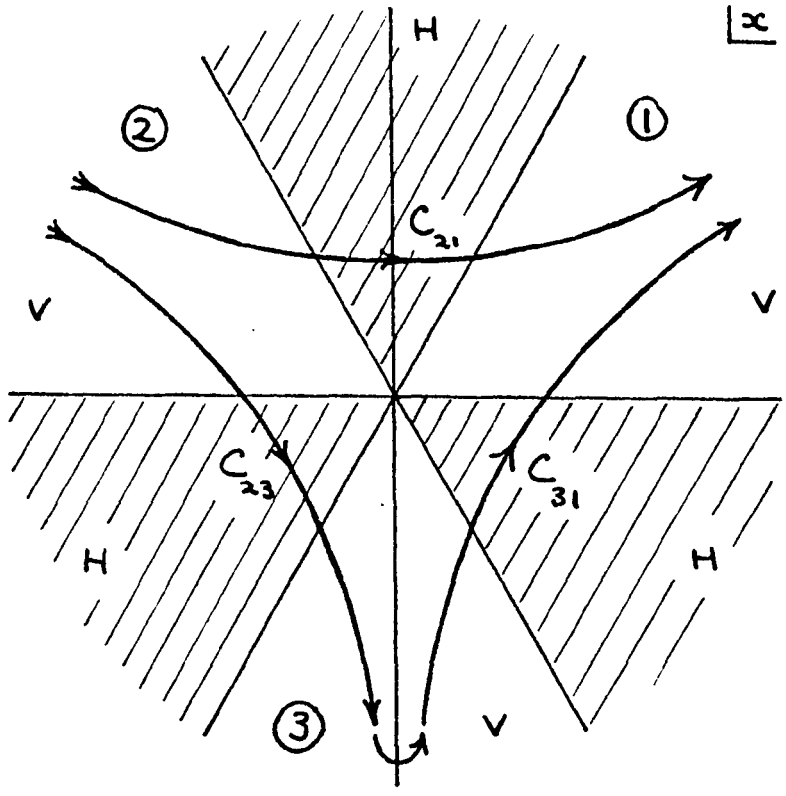


Figure 7.2 Integration Contour in the x-Plane

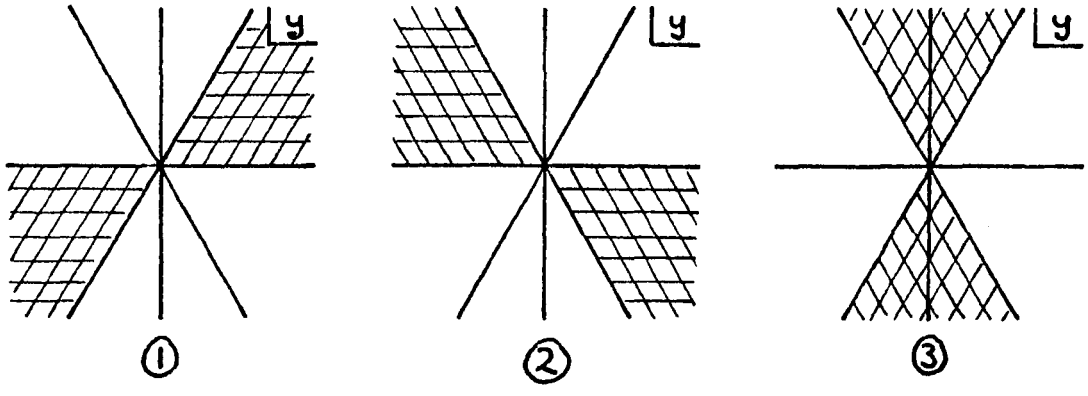


Figure 7.3 Sectors of the y-Plane

Now we can rotate the XY axes through $+\pi/4$ by applying the inverse of (7.7) to give

$$E(W_1,W_2,W_3) = \frac{e^{-i\pi/6}}{2^{1/3}} H(V_1,V_2,V_3) + \frac{e^{i\pi/6}}{2^{1/3}} H(V_1^*,V_2^*,V_3^*)$$

where $V_{\frac{1}{2}} = 2^{-2/3}(W_1 e^{i2\pi/3} \pm W_2 e^{i\pi/6}) = 2^{-2/3} e^{i2\pi/3}(W_1 \mp iW_2)$

and $V_3 = 2^{2/3} e^{i\pi/3} W_3$

If we define $Z = (W_1 + iW_2)/12^{1/3}$ and $U = (2/3)^{2/3} e^{i\pi/3} W_3$

then using (7.8) we have

$$E(W_1,W_2,W_3) = \left(\frac{2}{3}\right)^{2/3}\pi \sum_{n=0}^{\infty} \left\{ e^{-i\pi/6} \frac{(-iU)^n}{n!} Ai^{(n)}(-e^{i2\pi/3}Z^*)\, Ai^{(n)}(-e^{i2\pi/3}Z) \right. \quad (7.10)$$
$$\left. + e^{i\pi/6} \frac{(-iU^*)^n}{n!} Ai^{(n)}(-e^{-i2\pi/3}Z)\, Ai^{(n)}(-e^{-i2\pi/3}Z^*) \right\}$$

This series is useful for checking values of E for small W_3, where asymptotic stationary phase approximations are not accurate. In particular

$$E(W_1,W_2,0) = \left(\frac{2}{3}\right)^{2/3}\pi \{ e^{-i\pi/6} Ai(-e^{i2\pi/3}Z)\, Ai(-e^{i2\pi/3}Z^*)$$
$$+ e^{i\pi/6} Ai(-e^{-i2\pi/3}Z)\, Ai(-e^{-i2\pi/3}Z^*) \}$$

and using the formula

$$Ai(e^{\pm i2\pi/3}z) = \tfrac{1}{2} e^{\pm i\pi/3} \{ Ai(z) \mp iBi(z) \} \quad (7.11)$$

(see Abramowitz & Stegun, p446) this simplifies to

$$E(W_1,W_2,0) = (2/3)^{2/3}\pi \, Re\{ Ai(-Z)\, Bi(-Z^*) \} \quad (7.12)$$

which is the analogue of the result for H. This formula is also quoted without derivation by Trinkaus & Drepper (1977) who presumably used it in their computation of $E(W_1,W_2,0)$. It is only really suitable near the origin where the series for Ai and Bi converge rapidly, and Trinkaus & Drepper only compute such a small region.

It is not obvious that (7.12) has the required 3-fold symmetry. A rotation in the W_1W_2 plane is equivalent to the same rotation in the complex Z plane. Using (7.11) plus its analogue

$$Bi(e^{\pm i2\pi/3}z) = \tfrac{1}{2} e^{\mp i\pi/6} \{ 3Ai(z) \pm iBi(z) \}$$

one can show that

$$\{Ai(z)\ Bi^*(z) + Ai^*(z)\ Bi(z)\}/2$$

$$\equiv \{Ai(z)\ Bi^*(z) + Ai(e^{i2\pi/3}z)\ Bi^*(e^{i2\pi/3}z) + Ai(e^{-i2\pi/3}z)\ Bi^*(e^{-i2\pi/3}z)\}/3$$

The first form clearly shows the symmetry under $z \to z^*$, the second the rotation symmetry. In passing we notice that we have evaluated the non-trivial Fourier integral

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dx\ Ai(a - x^2)\ e^{ibx} = 2^{-1/3}\ Re\left\{Ai\left(\frac{a+ib}{2^{2/3}}\right)\ Bi\left(\frac{a-ib}{2^{2/3}}\right)\right\}$$

and its simpler companion

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dx\ Ai(a + x^2)\ e^{ibx} = 2^{-1/3}\ Ai\left(\frac{a+b}{2^{2/3}}\right)\ Ai\left(\frac{a-b}{2^{2/3}}\right)$$

Putting the asymptotic forms for the Airy functions in (7.11) gives a simple analytical expression for $E(W_1,W_2,0)$ well away from the origin. This shows an overall decrease of amplitude as $1/\sqrt{r}$ where r is the distance from the origin, and gives an explicit equation for the null lines, which agrees well with the computations. On the axis the series (7.10) gives

$$E(0,0,W_3) = (2/3)^{2/3}\sqrt{3}\pi\{(Ai(0))^2 - i(2/3)^{2/3}(Ai'(0))^2W_3 + O(W_3{}^3)\}$$

(Note that $W_3{}^2$ does not contribute). In particular

$$E(0,0,0) = (2/3)^{2/3}\sqrt{3}\pi(Ai(0))^2 = \frac{\Gamma(1/6)}{6\sqrt{\pi}} \simeq 0.5234$$

and $E(0,0,1) \simeq 0.56e^{-i0.12\pi}$

both of which agree well with computations, and show that the maximum amplitude does not occur at the origin.

7.2 Computational Method and Results

We notice that the highest power of y in (7.1) is quadratic, therefore we can perform the y integration analytically. After making the transformation $x \to x - W_3/3$ to simplify the coefficient of y^2 we have

$$E = \frac{e^{i(w_1-4W_3^2/9)W_3/3}}{2\pi} \int_{-\infty}^{\infty} dx \left\{ \int_{-\infty}^{\infty} dy\ e^{-i(3xy^2+W_2y)} \right\} e^{i\{x^3-2W_3x^2-(W_1-W_3^2)x\}}$$

Consideration of $|e^{-ixy^2}|$ (see fig.7.3) shows that we can rotate the y contour through $-\pi/4$ when $x > 0$ i.e. in sector (1) of fig.7.2, and through $+\pi/4$ when $x < 0$ i.e. in sector (2) of fig.7.2. We cannot make the same rotation of the y contour for all x, therefore we must split the x integral into two halves, and use

$$\int_{-\infty}^{\infty} dy\ e^{-i(3xy^2+W_2y)} = e^{i\{W_2^2/12x - \pi/4\ \mathrm{sgn}(x)\}} \sqrt{\pi/3|x|}$$

to give

$$E = \frac{e^{i(W_1-4W_3^2/9)W_3/3}}{2\sqrt{3\pi}} \left\{ e^{-i\pi/4} \int_0^{\infty} \frac{dx}{\sqrt{x}}\ e^{i\{x^3-2W_3x^2-(W_1-W_3^2)x+W_2^2/12x\}} + e^{+i\pi/4} \int_{-\infty}^{0} \frac{dx}{\sqrt{-x}}\ e^{i\{x^3-2W_3x^2-(W_1-W_3^2)x+W_2^2/12x\}} \right\} \quad (7.13)$$

This can also be written as

$$E = \frac{e^{i(W_1-4W_3^2/9)W_3/3}}{2\sqrt{3\pi}}\ e^{-i\pi/4} \int_{C_{21}} \frac{dx}{\sqrt{x}}\ e^{i\{x^3-2W_3x^2-(W_1-W_3^2)x+W_2^2/12x\}}$$

C_{21} distorted below branch cut.

(see fig.7.2) where x has a branch cut from the origin, somewhere in the **upper** half plane. For computation we use (7.13), and remove the integrable singularity at the origin by putting $x \to x^2$, $x \to -x^2$ respectively in the two integrals to give

$$E(W_1,W_2,W_3) = \frac{e^{i(W_1-4W_3^2/9)W_3/3}}{\sqrt{3\pi}} \left\{ \left[e^{-i\pi/4} \int_0^{\infty} dx\ e^{i\phi_+(x)} \right] + \left[e^{-i\pi/4} \int_0^{\infty} dx\ e^{i\phi_-(x)} \right]^* \right\} \quad (7.14)$$

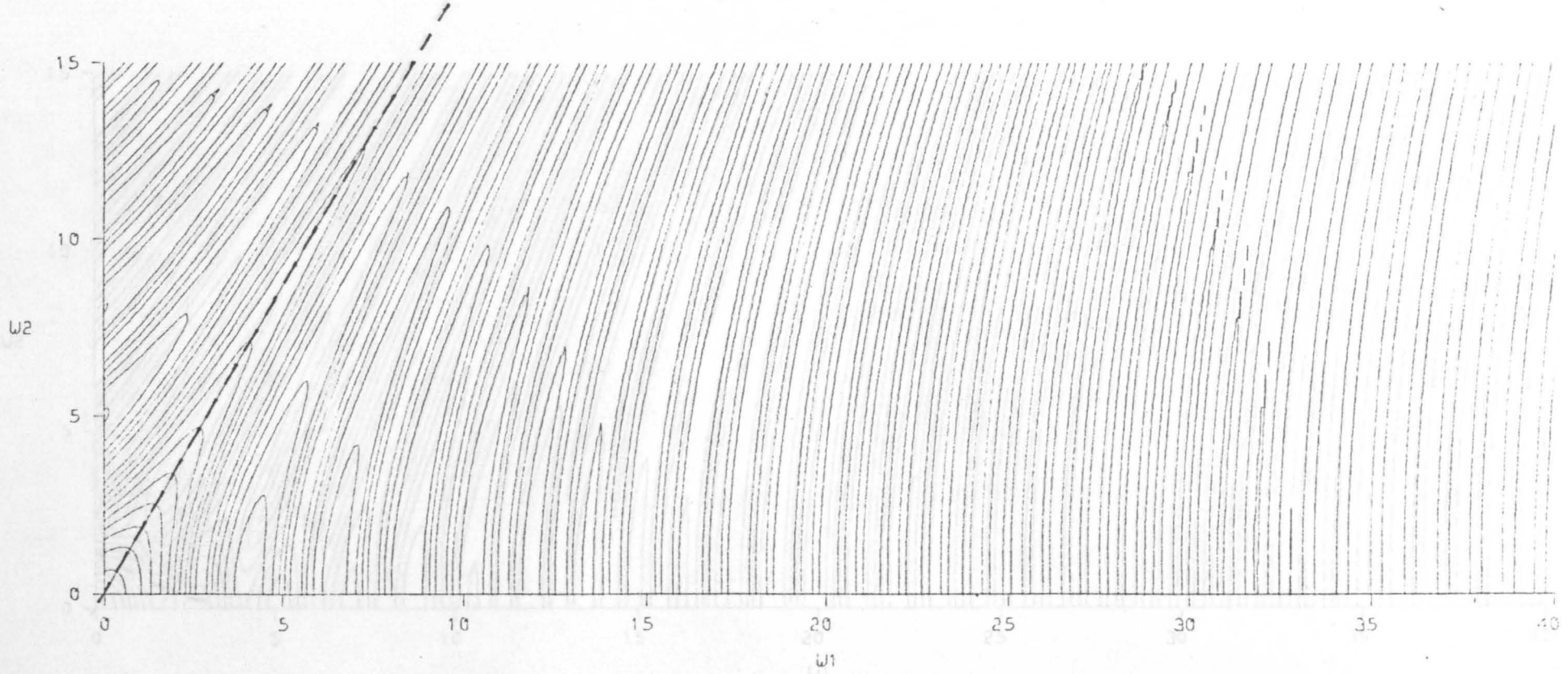
where $\phi_{\pm}(x) = x^6 \mp 2W_3x^4 - (W_1-W_3^2)x^2 + W_2^2/12x^2$ (7.15)

These two integrals are computed by the methods of chapter 4, and the details are discussed in appendix A7.1. We primarily compute sections W_3 = constant in the quadrant with W_1 & $W_2 \geq 0$. The 60° sector bounded by the W_1 axis is sufficient to generate the whole W_1W_2 pattern from the symmetry, leaving a redundant 30° sector. Because this redundant sector is computed independently it provides an excellent check on the computation.

Fig.7.4 shows contours of amplitude in the singular section $W_3 = 0$. The phase is not interesting since E is real. Figs.7.5 to 7.7 show contours of amplitude and phase for $W_3 = 2$, 4 & 6, and fig.7.8 shows simulations of the appearance of the optical diffraction pattern. These were produced by shading the contour plots so that black represents low intensity and white represents high intensity. The 60° sectors were then photographically reproduced, stuck together and rephotographed to create the whole pattern. These simulations are much more informative than attempts to label the contours.

The central region of fig.7.4 agrees with that displayed by Trinkaus & Drepper (1977). The caustic is a point at the origin, surrounded by a central triangular star shaped amplitude maximum. Around this are a sequence of fringes having maximum amplitude along the W_1 axis and its symmetry relatives, and the overall decrease of amplitude as $1/\sqrt{r}$ mentioned earlier is clearly visible. Along each dark fringe is an actual zero of amplitude.

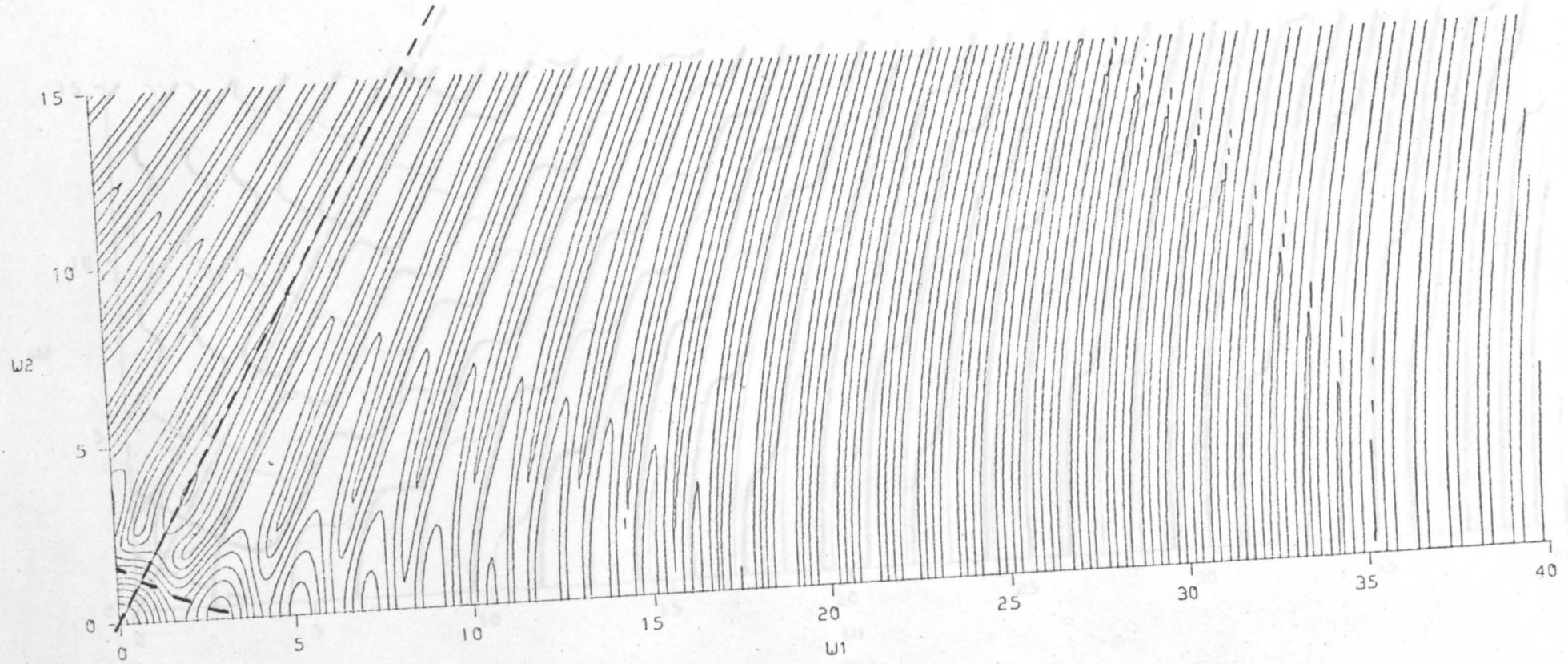
As W_3 increases the caustic expands and the pattern inside it breaks up into bright spots and becomes rapidly more complex. The phase plots show that the line nulls have gone leaving only isolated point nulls, where the thick and thin lines cross. The dark fringes outside the caustic no longer contain line zeros. The detailed structure of the elliptic umbilic diffraction pattern was first studied by Prof.J.F.Nye in an optical experiment. This is described by Berry, Nye & Wright (1978) who compare the experimental photographs with the computer simulations presented here. The most striking feature of the diffraction pattern is the dark hexagonal rings, clearly visible on the $W_3 = 4,5$ & 6 simulations. Among these rings are many groups of three black dots which look like the vestiges of hexagonal rings. The hexagons and vestiges of hexagons appear to lie on a close packed lattice, especially near to the centre. Observing this pattern unfolding under the microscope led Nye to postulate that the underlying structure of the elliptic umbilic diffraction pattern within the caustic was a hexagonal close packed lattice of puckered hexagonal null loops. He suggested that the lattice planes were slightly curved, to explain the change of structure away from the centre. Now it must be remembered that what one sees as dark regions on the experimental photographs and on the simulations are not points of zero amplitude, but regions where the amplitude is less than some particular small value (0.05 for the simulations). Therefore



ELLIPTIC UMBILIC [W3=0 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·05

Figure 7.4

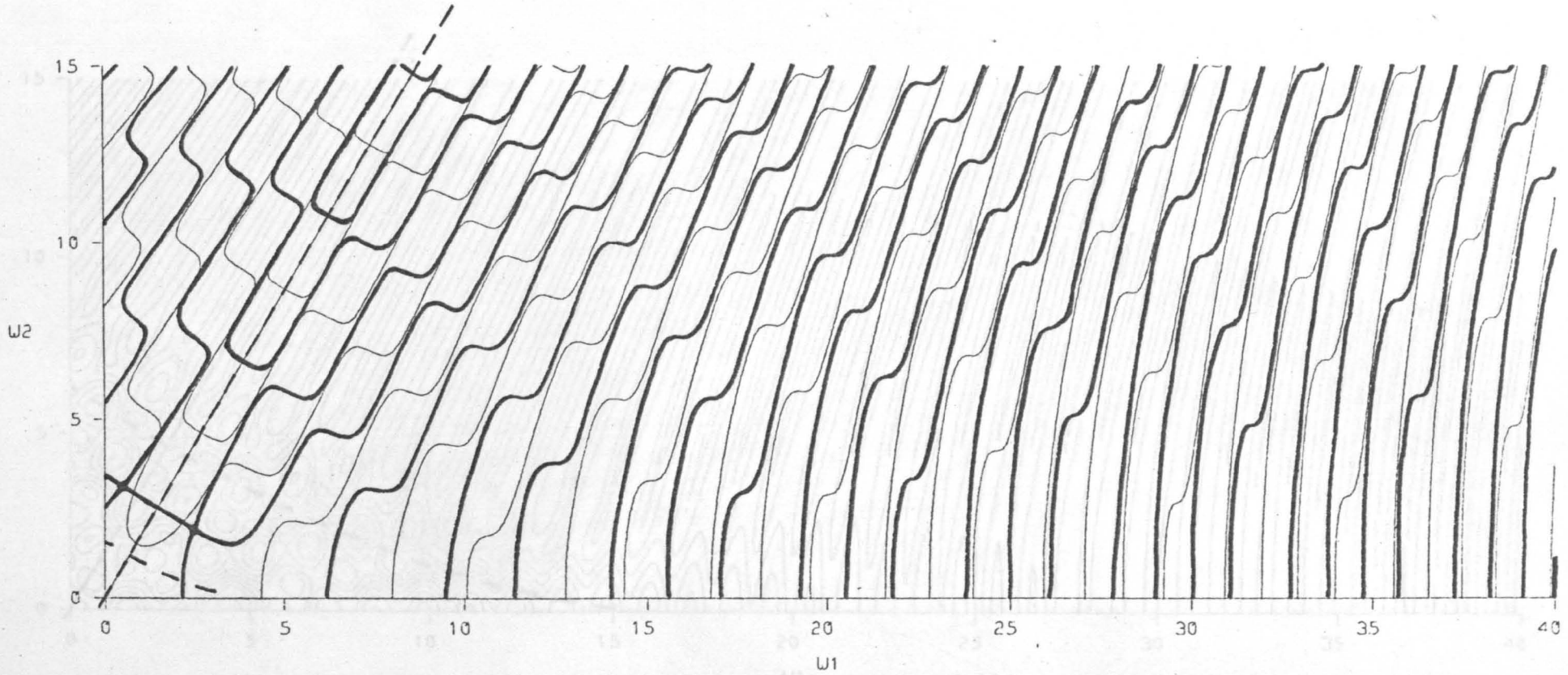
Maximum Amplitude = 0.52



ELLIPTIC UMBILIC [W3=2 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·05

Maximum Amplitude = 0.61

Figure 7.5a

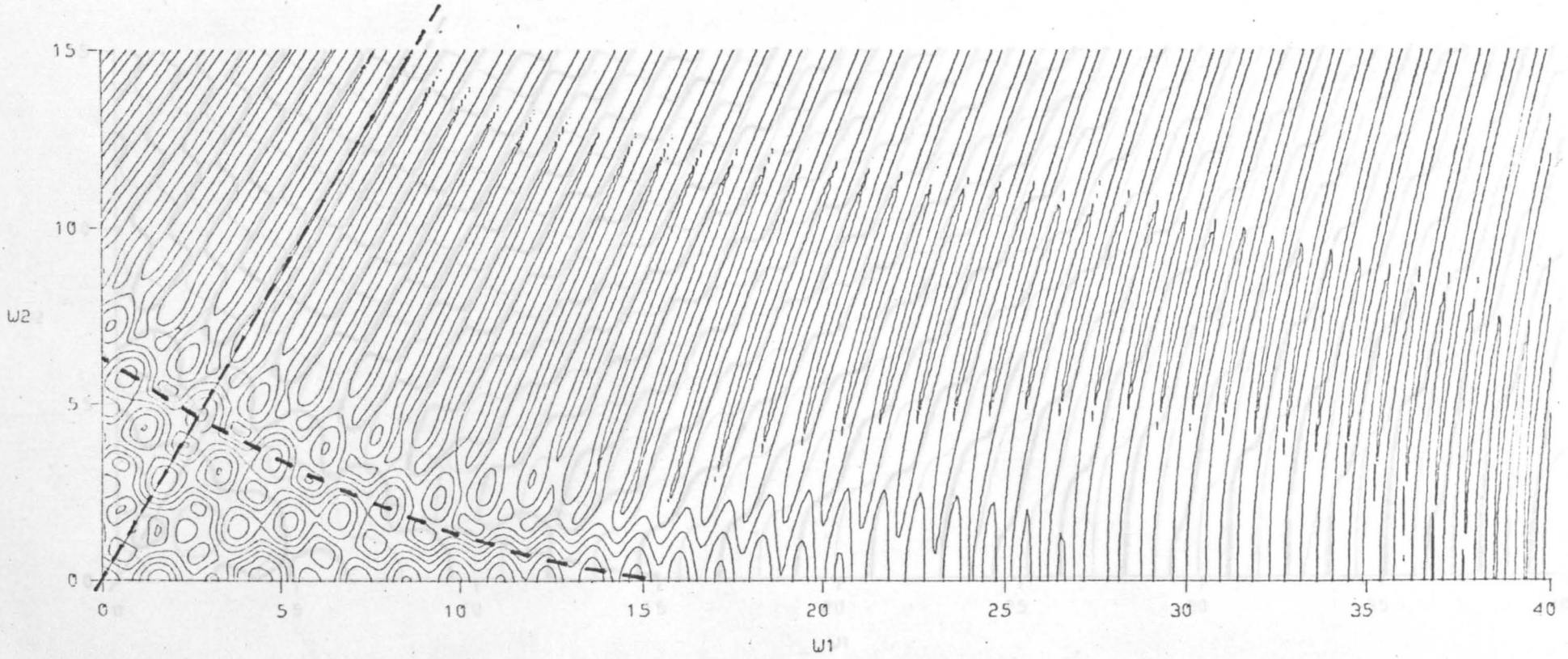


ELLIPTIC UMBILIC [W3=2 SECTION]
CONTOURS OF PHASE in steps of Pi/2

Thick lines 0,π

Thin lines ±π/2

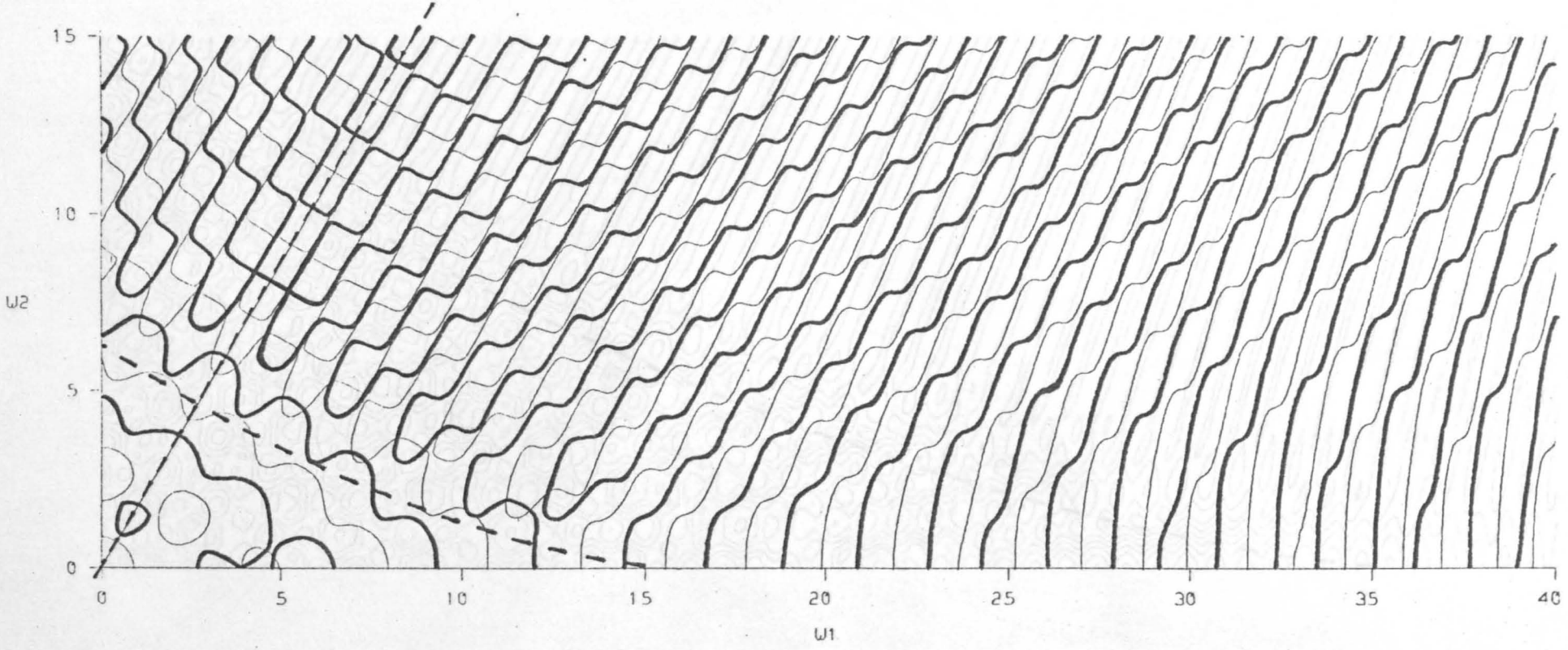
Figure 7.5b



ELLIPTIC UMBILIC [W3=4 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·05

Figure 7.6a

Maximum Amplitude = 0.42

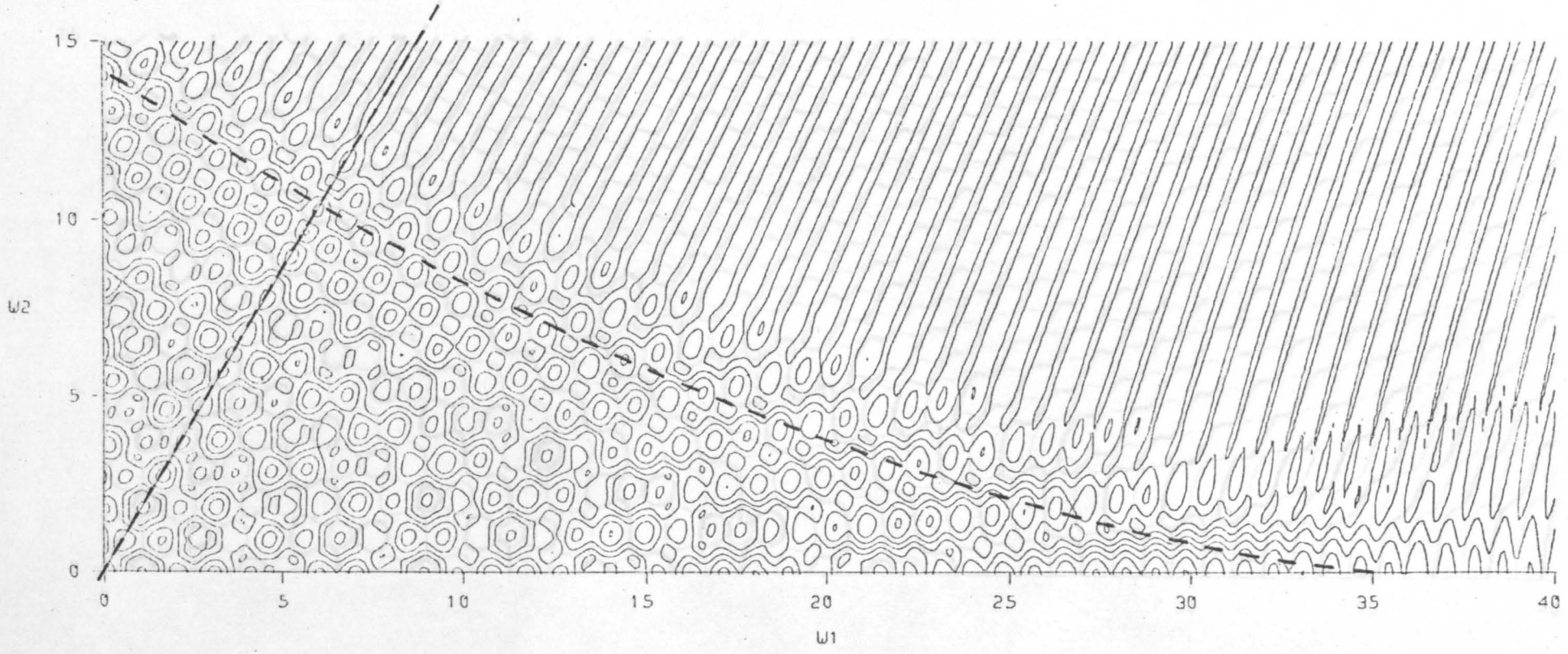


ELLIPTIC UMBILIC [W3=4 SECTION]
CONTOURS OF PHASE in steps of Pi/2

Thick lines 0,π

Thin lines ±π/2

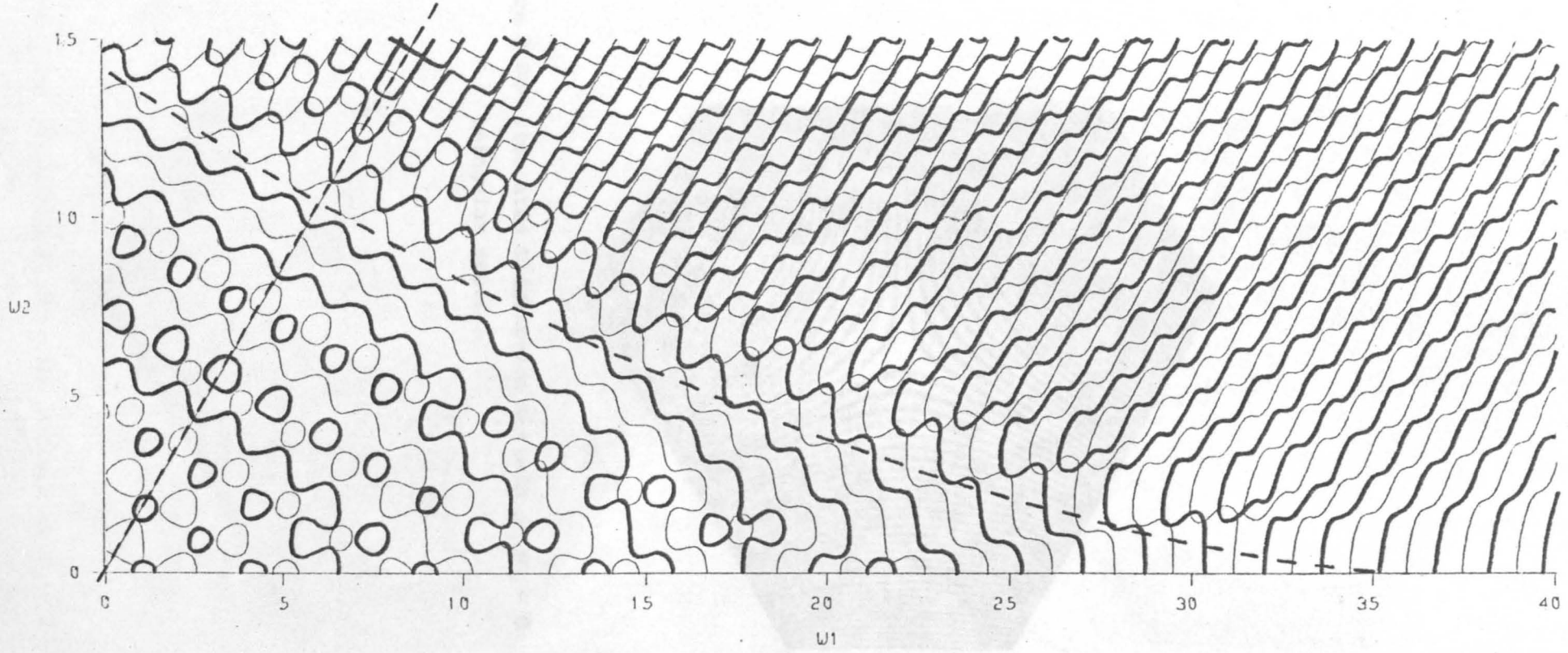
Figure 7.6b



ELLIPTIC UMBILIC [W3=6 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·05

Figure 7.7a

Maximum Amplitude = 0.36



ELLIPTIC UMBILIC [W3=6 SECTION]
CONTOURS OF PHASE in steps of Pi/2

Thick lines 0,π

Thin lines ±π/2

Figure 7.7b

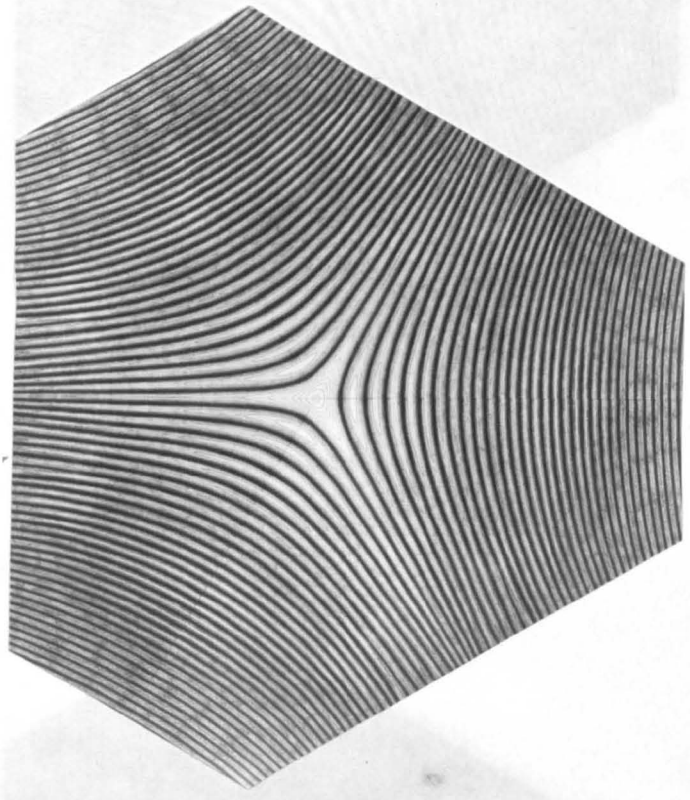


Figure 7.8a Simulated Diffraction Pattern at $W_3 = 0$ (singular section)

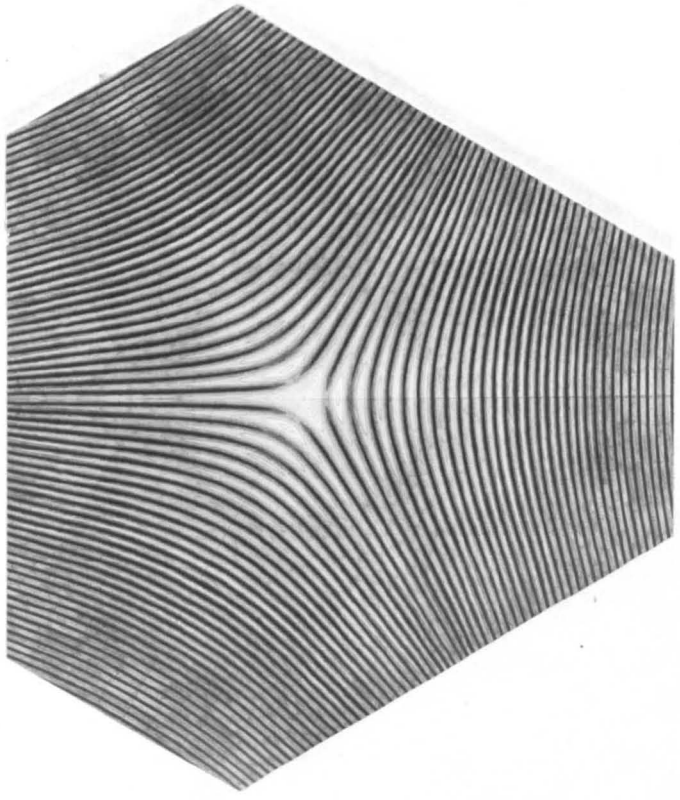


Figure 7.8b Simulated Diffraction Pattern at $W_3 = 1$

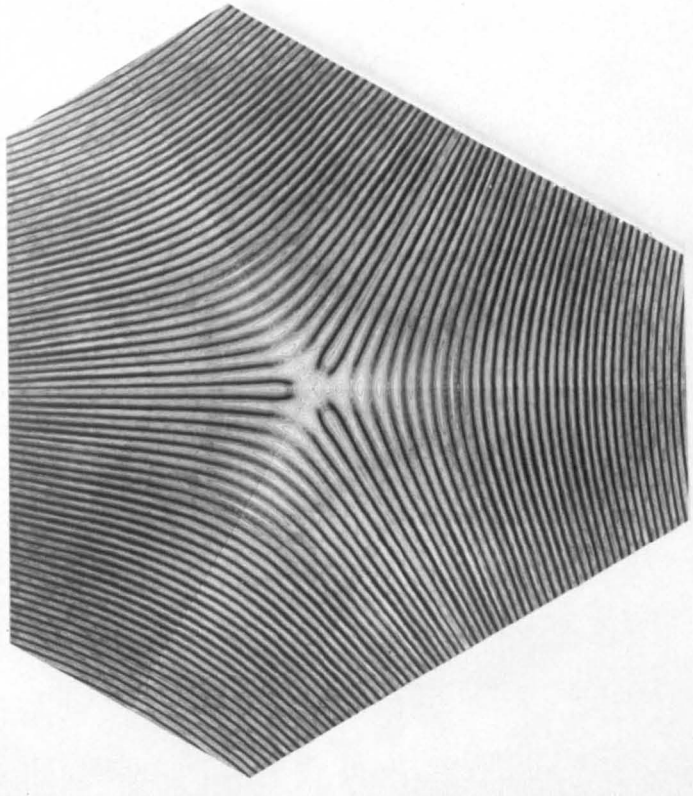


Figure 7.8c Simulated Diffraction Pattern at $W_3 = 2$

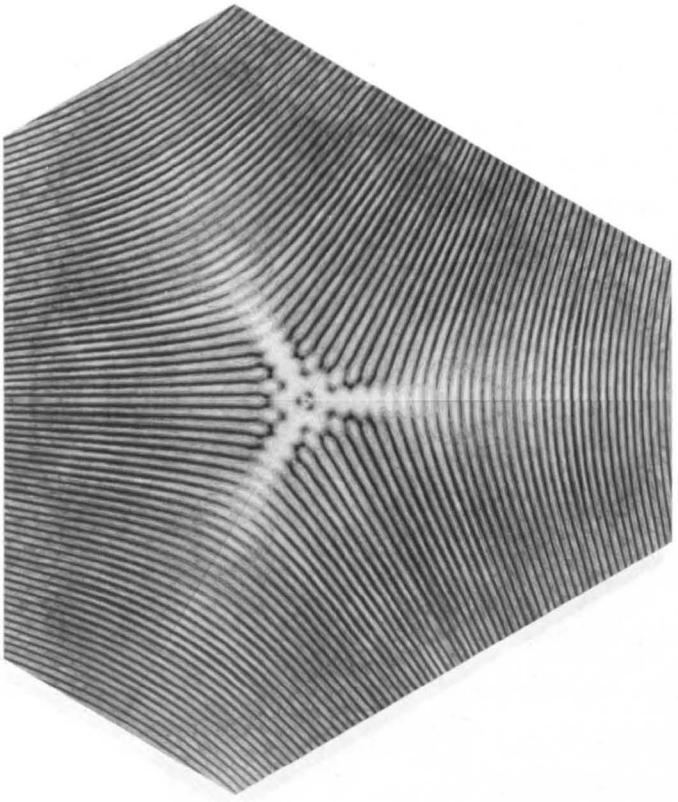


Figure 7.8d Simulated Diffraction Pattern at $W_3 = 3$

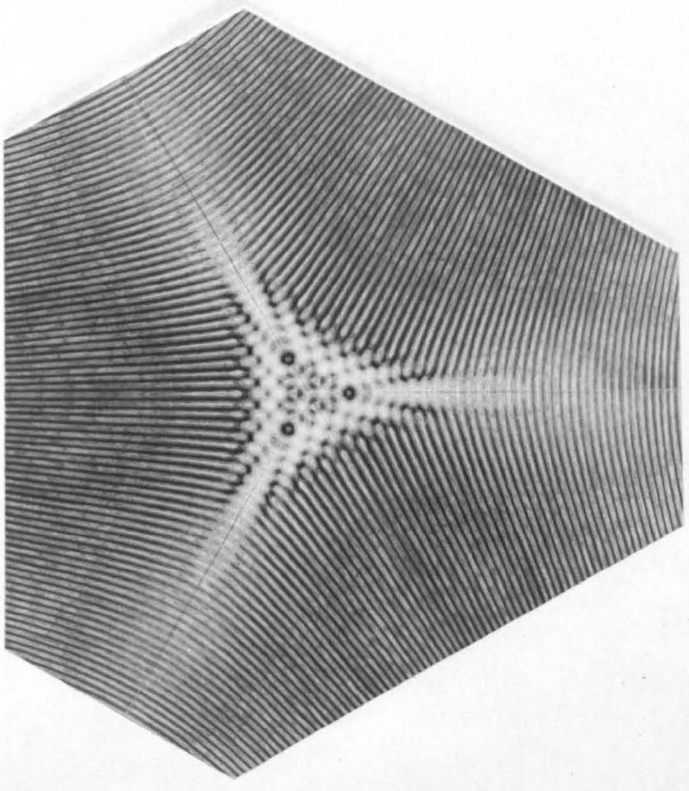


Figure 7.8e Simulated Diffraction Pattern at $W_3 = 4$

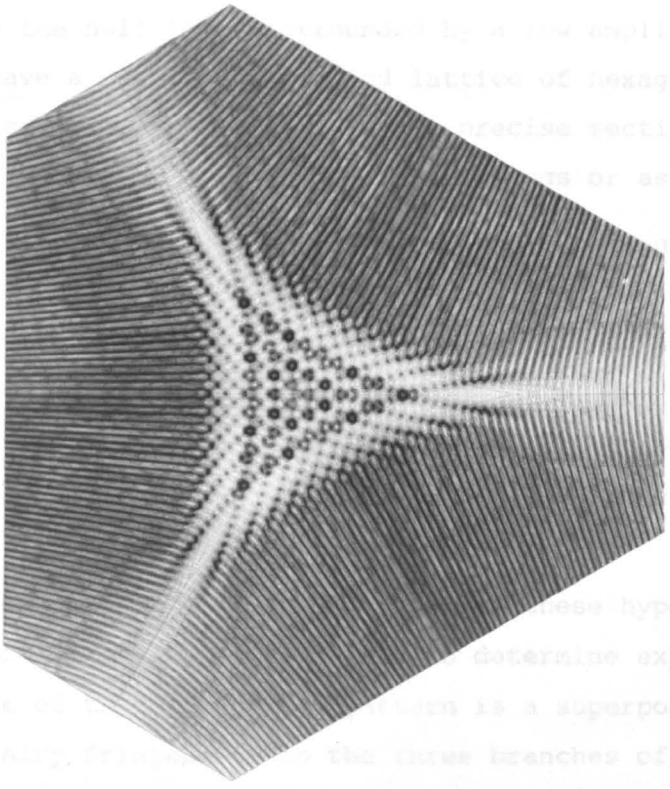


Figure 7.8f Simulated Diffraction Pattern at $W_3 = 5$

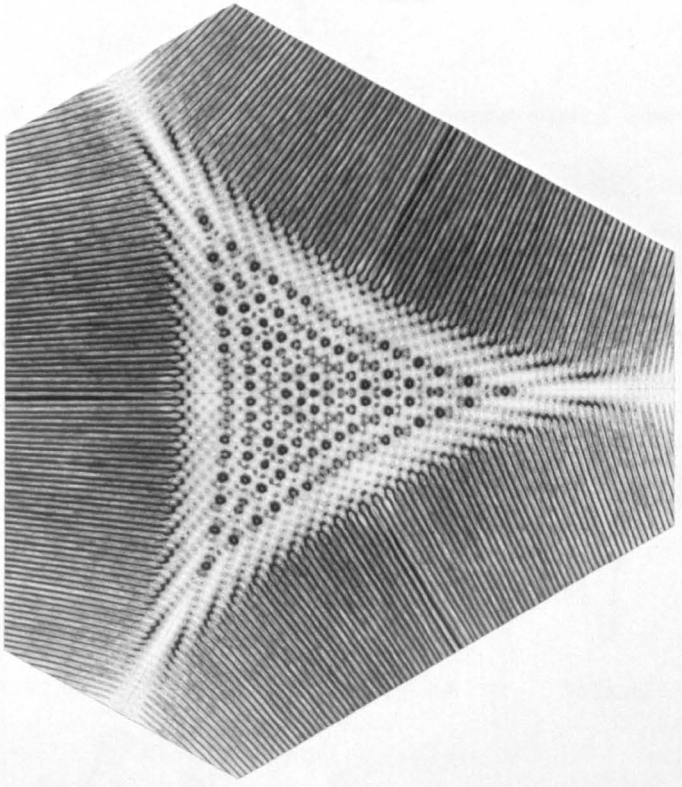


Figure 7.8g Simulated Diffraction Pattern at $W_3 = 6$

one should imagine the null lines surrounded by a low amplitude "black sausage". So we have a slightly distorted lattice of hexagonally puckered black sausages, and depending on the precise section one takes through these they may appear as full hexagonal rings or as three dots.

As we go to increasing W_3, the pattern in the cusps must approach the canonical cusp pattern, with its null pairs inside the caustic. Therefore Nye suggested that the distortion of the lattice becomes so large that neighbouring null loops link up into a "hairpin" whose arms are distorted helices. The amplitude of these helices would decrease with increasing W_3. Similarly we expect a single row of null lines running up outside the caustic for large W_3.

The following theoretical analysis confirms these hypotheses, and provides details of the null lines too fine to determine experimentally. The gross structure of the diffraction pattern is a superposition of the three sets of Airy fringes due to the three branches of the caustic, plus the closely spaced fringes running roughly perpendicular to the caustic. It is the effect of these closely spaced fringes which turn the triangular symmetry into approximate hexagonal symmetry near the centre. Nye's detailed analysis of this symmetry and other details are described by Berry, Nye & Wright (1978).

7.3 Null Lines inside the Caustic

We shall evaluate (7.1) by two-dimensional stationary phase using the formula

$$E \simeq \sum_i \frac{K_i \, e^{i\phi(x_i,y_i)}}{\sqrt{|H\phi(x_i,y_i)|}} \tag{7.16}$$

where (x_i,y_i) is a SPP satisfying

$$\nabla\phi(x_i,y_i) = 0$$

and the Hessian

$$H\phi = \phi_{xx}\,\phi_{yy} - (\phi_{xy})^2$$

$$K_i = 1 \quad \text{if } H < 0 \qquad \text{(saddle)}$$

$$+i \text{ if } H > 0 \text{ and } \phi_{xx} > 0 \text{ and } \phi_{yy} > 0 \qquad \text{(minimum)}$$

$$-i \text{ if } H > 0 \text{ and } \phi_{xx} < 0 \text{ and } \phi_{yy} < 0 \qquad \text{(maximum)}$$

The formula (7.16) is derived by expanding ϕ about its SPPs up to second order and diagonalizing the resulting quadratic form. For $W_2 = 0$ we can easily solve the ray equations (7.3) to give the SPPs:

$$x = \frac{W_3 \pm \sqrt{W_3^2 + 3W_1}}{3} \qquad y = 0$$

$$x = \frac{-W_3}{3} \qquad y = \pm\sqrt{\frac{W_3^2 - W_1}{3}} \tag{7.17}$$

We only consider $-W_3^2/3 < W_1 < W_3^2$, so the square roots are always real.

We consider the simplest case first, which is near the W_3 axis with W_1 and W_2 both $<<$ W_3. We use the SPPs (x_0,y_0) and Hessian as on the W_3 axis, and find the phase from

$$\phi(x,y,\underline{W}) = \phi(x_0,y_0,0,0,W_3) - W_1x_0 - W_2y_0$$

as described in chapter 5, to give

$$\begin{aligned} E(\underline{W}) &\simeq \frac{1}{2W_3}\left\{\frac{e^{-i4W_3^3/27}}{\sqrt{3}}\left[e^{-i2W_1W_3/3} + e^{i(W_1W_3/3 - W_2W_3/\sqrt{3})} + e^{i(W_1W_3/3 + W_2W_3/\sqrt{3})}\right] - i\right\} \\ &= \frac{1}{2W_3}\left\{\frac{e^{i(W_1 - 4W_3^2/9)W_3/3}}{\sqrt{3}}\left[e^{-iW_1W_3} + 2\cos\frac{W_2W_3}{\sqrt{3}}\right] - i\right\} \end{aligned} \tag{7.18}$$

This approximation still has the correct symmetry. Actually on the axis

$$|E(0,0,W_3)| \simeq \frac{1}{W_3}\sqrt{1 + \frac{\sqrt{3}}{2}\sin\frac{4}{27}W_3^3}$$

which has maxima and minima at

$$W_3 = \left[(2n \pm \tfrac{1}{2})27\pi/4\right]^{1/3}$$

respectively, but no nulls.

To analyze (7.18) further let us define

$$V_1 = W_1W_3, \quad V_2 = W_2W_3, \quad V_3 = 4W_3{}^3/27 \quad \text{and} \quad 2W_3E(\underline{W}) = F(\underline{V}).$$

Then

$$F(\underline{V}) = \frac{e^{i(V_1/3-V_3)}}{\sqrt{3}}\left[e^{-iV_1} + 2\cos\frac{V_2}{\sqrt{3}}\right] - i$$

We have effectively mapped our original wavefronts into equally spaced plane wavefronts. The function F not only has the rotation and reflection symmetry of E, but it also has translation symmetry with a period of $3\times2\pi$ in V_1, $\sqrt{3}\times2\pi$ in V_2 and 2π in V_3. We can therefore define lattice vectors $\underline{v}_1$, $\underline{v}_2$, $\underline{v}_3$ respectively. Then F is also invariant under a translation of $(\underline{v}_1 + \underline{v}_2)/2$ and $\pm(\underline{v}_1 + \underline{v}_3)/3$. Therefore the "plane wave-function" F has the translation symmetry of a cubic close packed lattice. The space-group symmetry may be represented by placing a triangular motif, with no mirror plane perpendicular to the W_3 axis, at each lattice point (e.g. fig.7.10, where the + and - have no significance at this stage other than that of being different).

The null lines F = 0 satisfy

$$\left.\begin{aligned} \cos V_1 + 2\cos V_2/\sqrt{3} &= \sqrt{3}\,\sin(V_1/3 - V_3) \\ -\sin V_1 &= \sqrt{3}\,\cos(V_1/3 - V_3) \end{aligned}\right\} \qquad \begin{aligned} &(7.19a) \\ &(7.19b) \end{aligned}$$

The solution of these equations has the symmetry of F plus rotation diads along $V_1 = 0$, $V_3 = \pm\pi/2$ and their symmetry relatives (these diads take F into -F*). Equation (7.19b) shows that a necessary condition on the null lines is that their projections onto the V_1V_3 plane lie on the curves

$$V_3 = V_1/3 - \arccos(-(\sin V_1)/\sqrt{3}) \qquad (7.20)$$

This curve must lie within isolated bands, and is easily sketched as in fig.7.9. Similarly we can find a necessary condition on the projection onto the V_1V_2 plane by squaring and adding (7.19a) and (7.19b) to give

$$V_1 = \arccos\left(\frac{-\cos 2V_2/\sqrt{3}}{2\cos V_2/\sqrt{3}}\right) \qquad (7.21)$$

This curve again lies within isolated bands, at the edges of which

$$\cos V_2/\sqrt{3} = \pm(-1 + \sqrt{3})/2 \qquad (7.22)$$

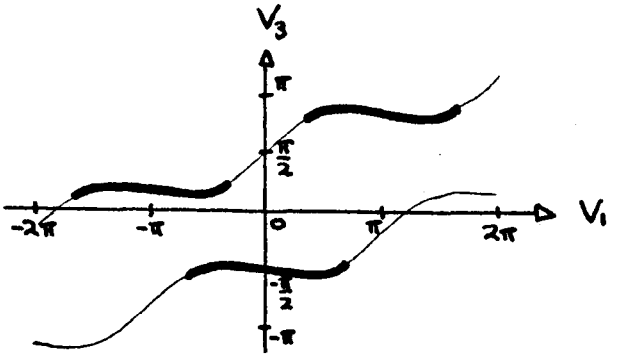


Figure 7.9 V_1V_3 Projection of Null Loops near Centre

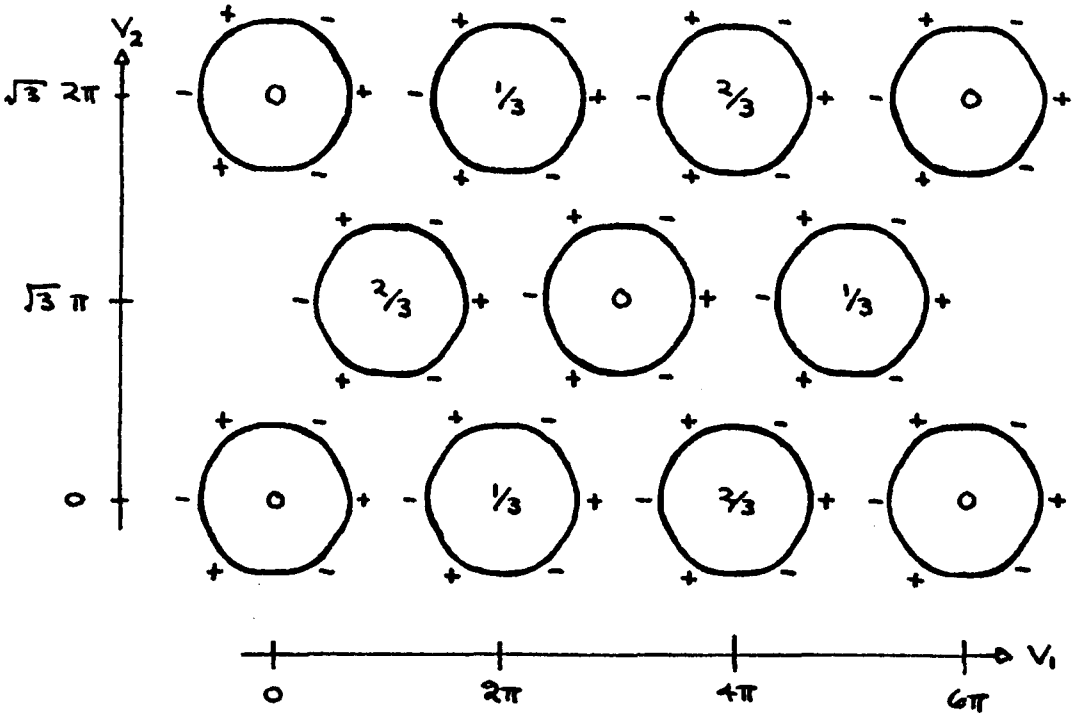


Figure 7.10 V_1V_2 Projection of Null Loops near Centre

Using (7.21) and its known symmetry we can easily sketch the V_1V_2 projection in fig.7.10. We expect hexagonal loops, but in fact they look more like circles. This is because the radius of the loop along V_1 is $2\pi/3 = 2.09$, and along V_2 is 2.07 from (7.22). These radii only differ by 1%, therefore the loops project into circles with a very slight hexagonal distortion which is hardly observable.

The loops can only have the required symmetry if they are centred on portions of fig.7.9 having negative slope, as indicated by the thicker line. Then the heights at the ends of the V_1V_3 projection are the same as the heights of the maximum and minimum, since $\sin(\pm 2\pi/3) = \sin(\pm(\cos\pi/3)2\pi/3)$. This agrees with establishing the correspondence of the different branches of (7.20) and (7.21) by substituting back in (7.19). Eqn.(7.20) gives the total depth of the loops as $\pi/9$, so that their depth/spacing = 1/18 and they do not occupy a great deal of space. On fig.7.10 we indicate the whole configuration by showing the fractional heights of the centres of the loops within the unit cell, and the puckering by ± signs to indicate heights of $\pm\pi/18$ relative to the centre of the loop.

The shape of the loops is given approximately in circular polar coordinates (r,θ) by the equations

$$r = 2\pi/3, \qquad V_3 = -\pi/2 - (\pi/18)\sin 3\theta$$

as shown in fig.7.12, where $V_1 = r \sin\theta$. Then

$$\begin{aligned} V_3 &= -\pi/2 - \pi\{3\sin\theta - 4\sin^3\theta\}/18 \\ &= -\pi/2 - V_1/4 + 3V_1^3/4\pi^2 \qquad \text{exactly} \\ &= -\pi/2 - 0.250\, V_1 + 0.076\, V_1^3 \end{aligned}$$

This is a "best" cubic approximation to (7.20), since it is exact at 3 points, effectively. In fact, the expansion of (7.20) up to third order is

$$V_3 = -\pi/2 - 0.244\, V_1 + 0.064\, V_1^3 \ .$$

The V_1V_3 projection of the null loops is a 3::1 Lissajous figure!

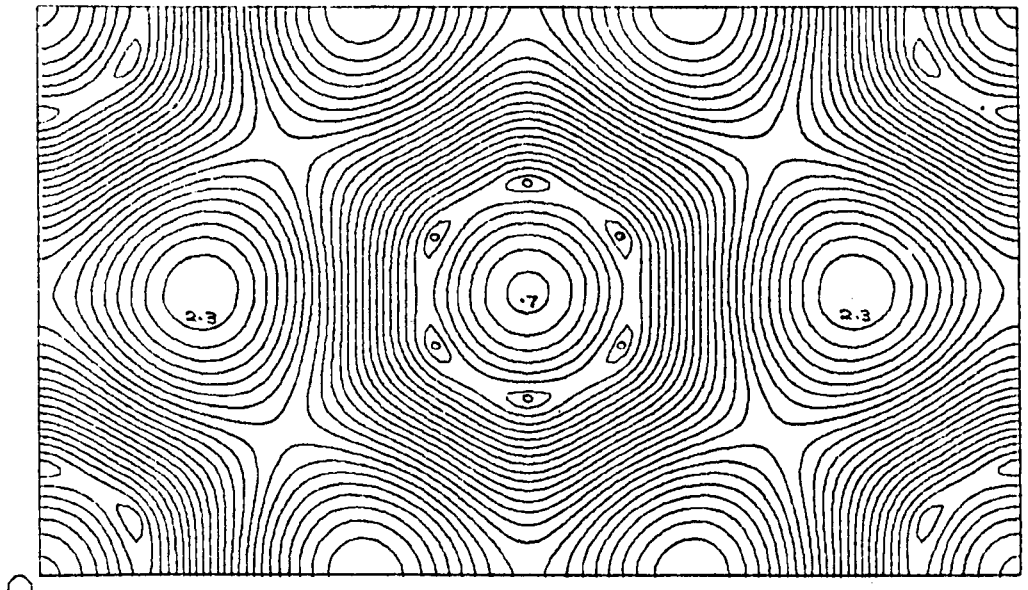
The null loops encircling the W_3 axis, one of which we take as the origin of our lattice, occur at $V_3 = 2n\pi - \pi/2$. We saw earlier that axial maxima/minima of amplitude occur at $V_3 = 2n\pi \pm \pi/2$. Therefore along the centre of any column of null loops the amplitude is minimal at the centre of a loop and maximal half way between loops. When we map our approximate solution back to real $\underline{W}$ space we lose all the translation

symmetry. The spacing of the lattice planes in which the null lines lie (the null planes) decreases as $1/W_3^2$ and the scale of the pattern within such a plane decreases as $1/W_3$, so the translation symmetry within these planes remains. However, the diffraction pattern is well described in terms of the distorted lattice in real space, and near the W_3 axis for large W_3 the lattice is locally not much distorted.

The hexagonal appearance of the black sausages is due to the hexagonal symmetry of their stacking, and the fact that the interstices in the lattice are filled by bright blobs which "push in" the sides of the black sausages. This effect is clearly shown in figs.7.11a,b,c, where we plot contours of some interesting sections of $F(\underline{V})$ in the same unit cell as shown in fig.7.10. Fig.7.11a is a plane through the middle of a null plane. The contour at amplitude 0.1 shows a set of 6 black spots, but the contour at 0.2 shows a hexagonal ring. The contours inside this ring are nearly circular, but those outside are constrained by the surrounding 6 bright spots to adopt a hexagonal shape, point-to-point with neighbouring hexagons in the lattice. Fig.7.11b is a plane touching the top of a null loop to give 3 isolated dark spots. The amplitude contours do not link up into a hexagonal ring until 0.4. Fig.7.11c is a plane half way between two sets of null loops, showing clearly the large hexagonal "interstitial" maximum of amplitude. Vestiges of the null loops above and below the plane remain at the points where they are closest together, as the set of low amplitude spots with a rectangular shape. They do <u>not</u> contain a zero.

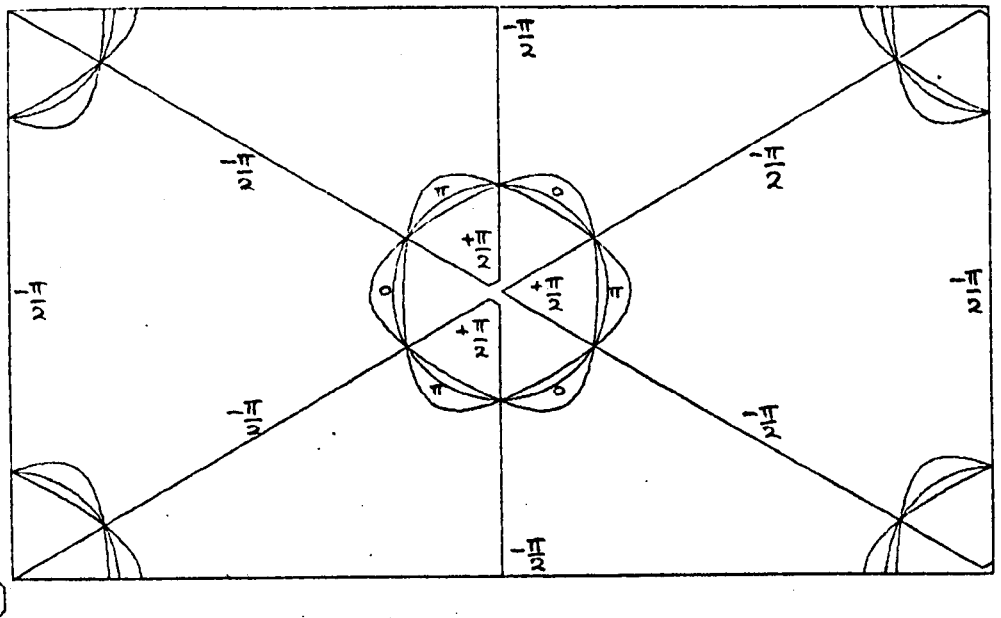
Armed with this information, we can check the agreement of our "axial" theory of puckered null loops with the actual computed diffraction patterns. We calculate where in a unit V_3 repeat the planes $W_3 = 3,4,5,6$ occur, and plot this along with the positions of the sets of null loops schematically in fig.7.13a. Then in fig.7.13b we sketch what we expect the darkest regions of the central diffraction patterns to look like, mapped back into $\underline{W}$ space. $W_3 = 3,4$ & 5 give us transition patterns similar to fig.7.11c, and $W_3 = 6$ gives full black rings as in fig.7.11a. These sketches compare very well with the simulations in fig.7.8 near the centre, but the agreement deteriorates away from the centre due to the curvature of the "real" lattice planes.

What this "axial" approximation does not tell us is how many null loops there are in any particular "plane". To find this we can expand about the fold caustic at $W_1 = -W_3^2/3$, $W_2 = 0$ by putting $W_1 = -W_3^2/3 + \delta$.



0

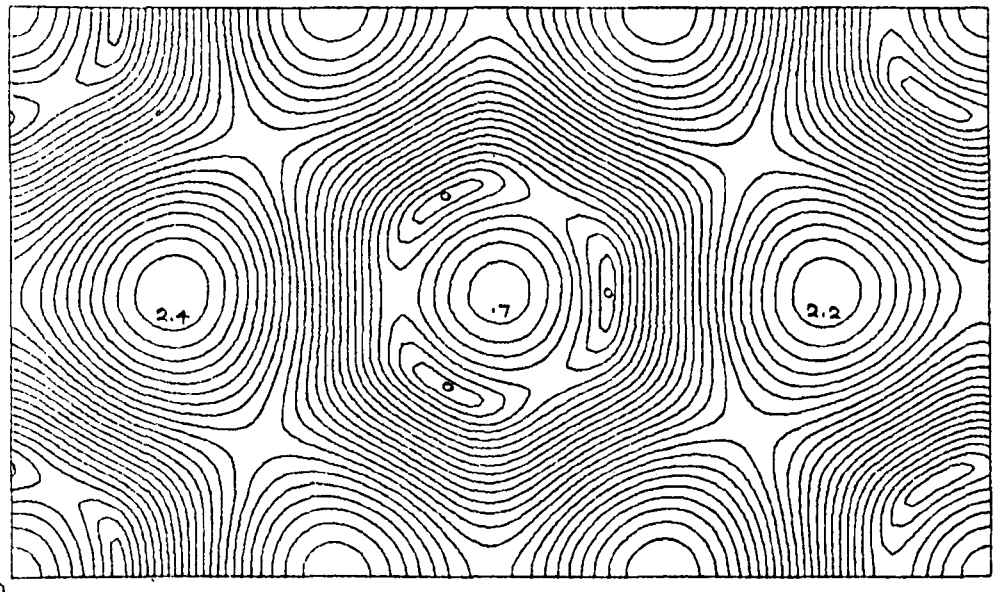
CONTOURS OF AMPLITUDE in steps of 0·1
AT FRACTIONAL HEIGHT 0 IN UNIT CEL



0

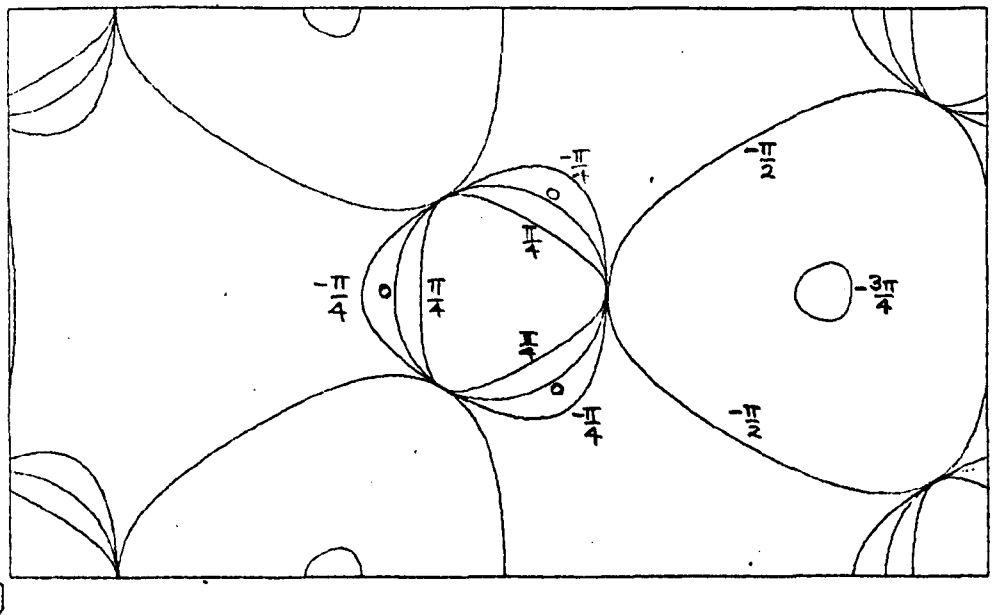
CONTOURS OF PHASE in steps of Pi/4
AT FRACTIONAL HEIGHT 0 IN UNIT CEL

Figure 7.11a



0

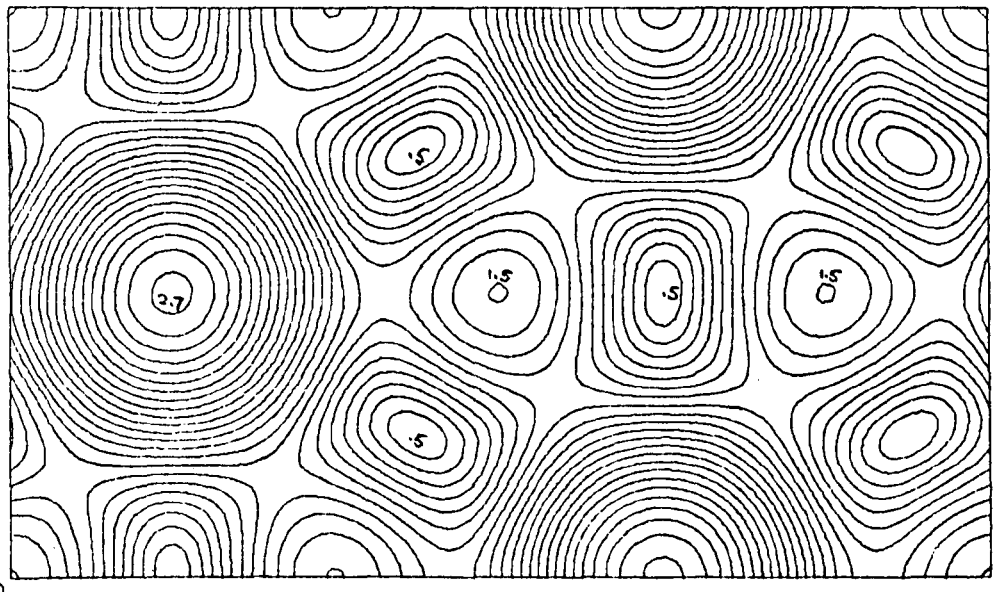
CONTOURS OF AMPLITUDE in steps of 0·1 AT FRACTIONAL HEIGHT 1/36 IN UNIT CELL



0

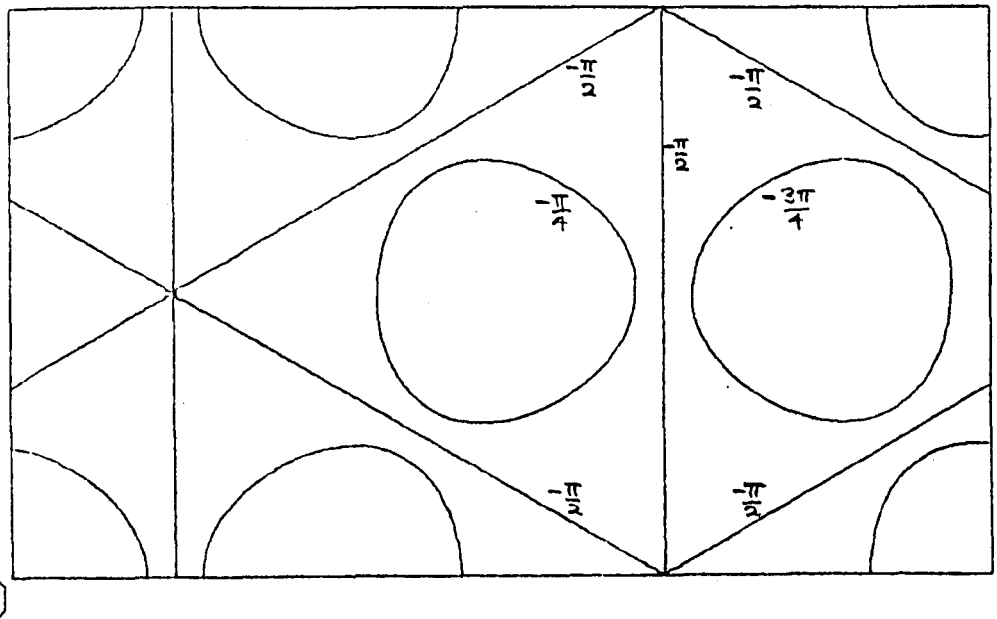
CONTOURS OF PHASE in steps of Pi/4 AT FRACTIONAL HEIGHT 1/36 IN UNIT CELL

Figure 7.11b



0

CONTOURS OF AMPLITUDE in steps of 0·1
AT FRACTIONAL HEIGHT 1/6 IN UNIT CELL



0

CONTOURS OF PHASE in steps of Pi/4
AT FRACTIONAL HEIGHT 1/6 IN UNIT CELL

Figure 7.11c

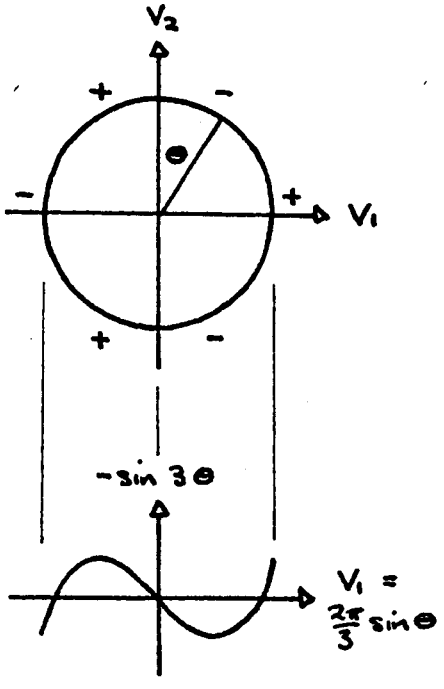


Figure 7.12 Central Null Loop

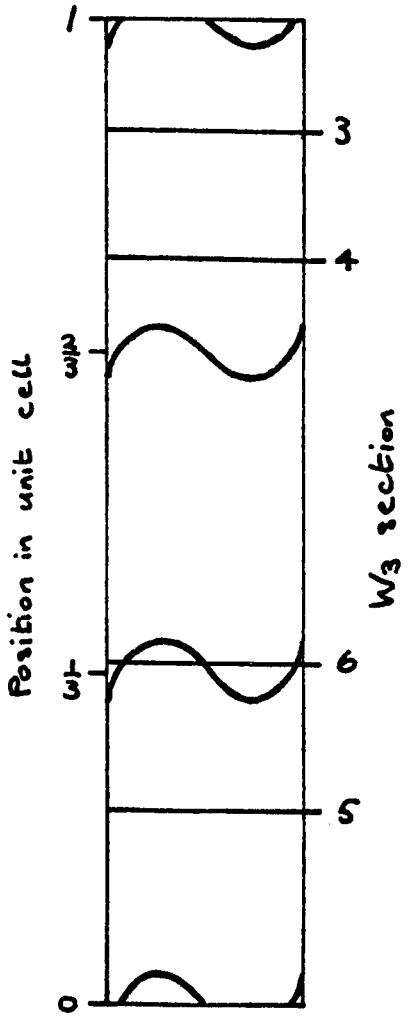


Figure 7.13a Positions of W_3 Planes in Unit Cell

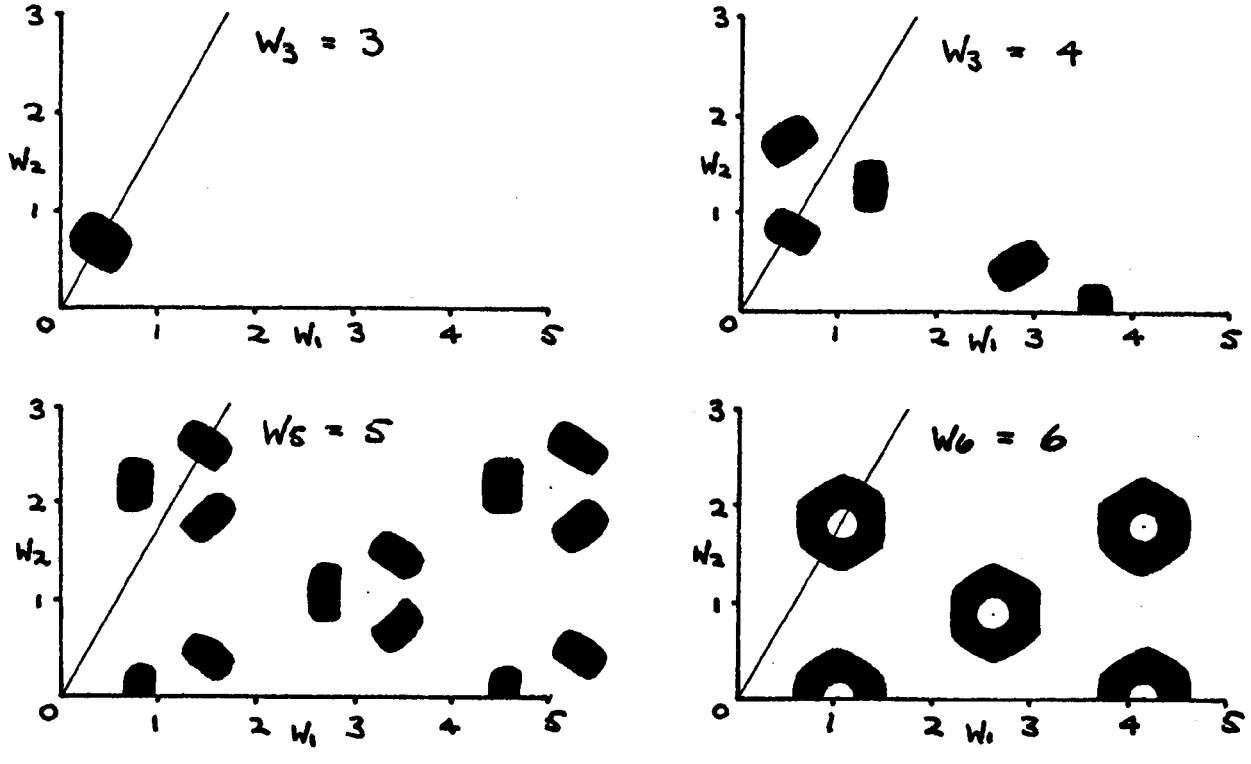


Figure 7.13b Central Regions of Diffraction Patterns

Then our standard methods give

$$E \simeq \frac{1}{2W_3}\left\{\cos(2W_2W_3/3)\,e^{i(\delta - 7W_3^2/9)W_3/3} + \sqrt{\frac{2W_3}{\sqrt{3}\delta}}\sin(\pi/4 + 2(\delta/3)^{3/2})\,e^{i\{(W_3^2/9 - \delta)W_3/3 - \pi/4\}}\right\}$$

The first term represents the closely spaced fringes perpendicular to the caustic, which pass straight through the caustic. The second term represents the wider Airy fringes inside the caustic, and does not contribute outside. The first term perturbs the Airy fringes and breaks them up into null loops. Then we expect the rows of loops to be centred on

$$\sin(\pi/4 + 2(\delta/3)^{3/2}) = 0 \quad \Rightarrow \quad \delta = 3\{(N - 1/4)\pi/2\}^{2/3}$$

The centre of the pattern occurs at $\delta = W_3^2/3$, therefore the number of rows of null loops up to and including the centre is

$$N = \mathrm{Int}\{1/4 + (W_3/3)^3\,2/\pi\}$$

A dark Airy fringe crosses the centre if

$$W_3 = 3\{(N - 1/4)\pi/2\}^{1/3}$$

i.e. $V_3 = 4W_3^3/27 = 2N\pi - \pi/2$

which is exactly the condition for a null loop to encircle the W_3 axis, from our axial approximation. This gives us confidence that we can reliably relate the number of rows of hexagons to W_3, as illustrated in fig.7.14. Although it is difficult to count the loops because of the curvature of the lattice, it is quite easy to count the rows of loops (i.e. Airy fringes), and by comparison with fig.7.14 estimate the W_3 value of particular diffraction patterns. This works reliably on the simulations in fig.7.8.

In order to investigate the distortions of the lattice of null loops we need a more accurate theory. However, this means that we are no longer able to solve the problem analytically. We have already, in eqn.(7.17), found the SPPs exactly everywhere in the plane $W_2 = 0$, and we can easily write down the exact stationary phase approximation to E in this case. For small non-zero W_2 we can then perturb this solution to give

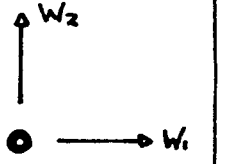
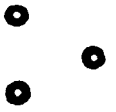
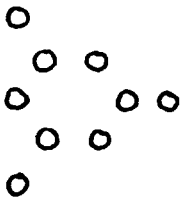
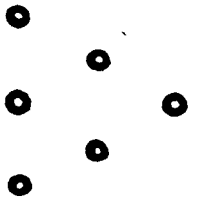
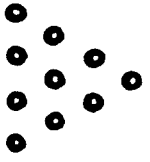
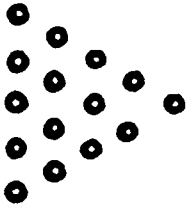
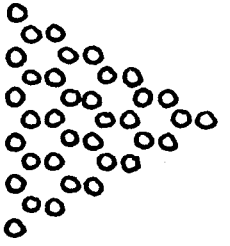
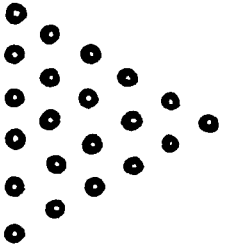
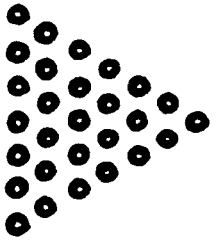
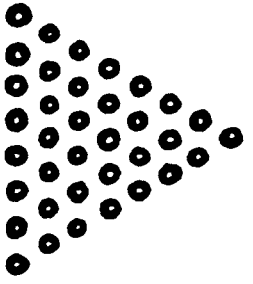
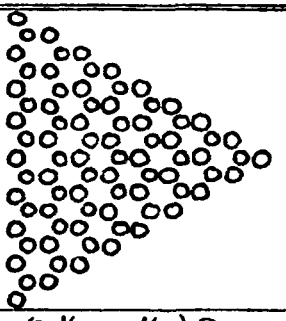
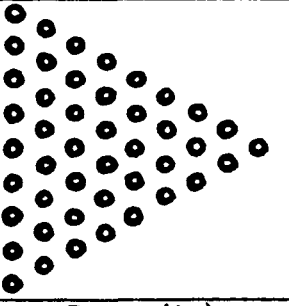
fractional height in unit cell	0	1/3	1/2	2/3
central intensity	minimum		maximum	
$N = 1$			vestiges of 1/3 & 2/3 	
V_3	$(1 - 1/4)2\pi$	$(1\,1/3 - 1/4)2\pi$	$(1\,1/2 - 1/4)2\pi$	$(1\,2/3 - 1/4)2\pi$
W_3	3.17	3.58	3.76	3.92
$N = 2$				
V_3	$(2 - 1/4)2\pi$	$(2\,1/3 - 1/4)2\pi$	$(2\,1/2 - 1/4)2\pi$	$(2\,2/3 - 1/4)2\pi$
W_3	4.20	4.45	4.57	4.68
$N = 3$				
V_3	$(3 - 1/4)2\pi$	$(3\,1/3 - 1/4)2\pi$	$(3\,1/2 - 1/4)2\pi$	$(3\,2/3 - 1/4)2\pi$
W_3	4.89	5.08	5.17	5.25

Figure 7.14 Triangle Numbers of Null Loops

$$E \simeq \frac{e^{i(-2W_3^3/27 - W_1W_3/3)}}{2}\left\{\frac{e^{-i2(W_3^2+3W_1)^{3/2}/27}}{\sqrt{(W_3^2+3W_1)+2W_3\sqrt{W_3^2+3W_1}}} - \frac{ie^{i2(W_3^2+3W_1)^{3/2}/27}}{\sqrt{-(W_3^2+3W_1)+2W_3\sqrt{W_3^2+3W_1}}}\right\}$$

$$+ e^{i(-4W_3^3/27 + W_1W_3/3)} \frac{\cos W_2\sqrt{(W_3^2 - W_1)/3}}{\sqrt{3(W_3^2 - W_1)}}$$

where two rays combine to give the cosine term. When expanded about the centre and the fold caustic this simplifies to the two previous approximate forms. Berry has found the exact zeros of the above expression in the case $W_2 = 0$. We generalize his method to $W_2 \neq 0$, by defining new variables

$$\beta = \frac{\pm 1\sqrt{W_3^2+3W_1}}{W_3}, \qquad \zeta = W_3^3/9 \quad \& \quad \xi = W_2\sqrt{\frac{W_3^2-W_1}{3}}$$

to give

$$E \simeq \frac{e^{i\{-2\zeta/3 - \zeta(\beta^2-1)\}}}{2W_3\sqrt{\beta}}\left\{\frac{e^{-i2\zeta\beta^3/3}}{\sqrt{2+\beta}} - ie^{\frac{i2\zeta\beta^3/3}{\sqrt{2-\beta}}}\right\}$$

$$+ e^{i\{-4\zeta/3 + \zeta(\beta^2-1)\}} \frac{\cos\xi}{W_3\sqrt{2-\beta}\sqrt{2+\beta}}$$

Since $W_1 = W_3^2(\beta^2-1)/3$, then for $W_2 = 0$ the fold occurs at $\beta = 0$, the centre at $\beta = 1$ and the cusp at $\beta = 2$. The condition for $E = 0$ is

$$\sqrt{2-\beta}\, e^{-i2\zeta\beta^3/3} - i\sqrt{2+\beta}\, e^{+i2\zeta\beta^3/3} + 2\sqrt{\beta}\cos\xi\, e^{i\zeta(-2/3 + 2\beta^2 - 2)} = 0$$

or $$2\sqrt{\beta}\cos\xi\, e^{i\zeta(2\beta^2 - 8/3 - 2\beta^3/3)} = i\sqrt{2+\beta} - \sqrt{2-\beta}\, e^{-i4\zeta\beta^3/3}$$

$$\Rightarrow \quad 2\sqrt{\beta}\cos\xi\cos\zeta(2\beta^3/3 - 2\beta^2 + 8/3) = -\sqrt{2-\beta}\cos 4\zeta\beta^3/3 \quad \text{(a)}$$

$$\text{and} \quad 2\sqrt{\beta}\cos\xi\sin\zeta(2\beta^3/3 - 2\beta^2 + 8/3) = -\sqrt{2+\beta} - \sqrt{2-\beta}\sin 4\zeta\beta^3/3 \quad \text{(b)} \qquad (7.23)$$

Squaring and adding gives

$$\sin 4\zeta\beta^3/3 = 2(\beta\cos\xi - 1)/\sqrt{4-\beta^2} \qquad (7.24)$$

and substituting this in (7.23b) gives

$$\sin\zeta(2\beta^3/3 - 2\beta^2 + 8/3) = -\sqrt{\frac{\beta}{2+\beta}}\left(\frac{\frac{1}{2} + \cos^2\xi}{\cos\xi}\right) \qquad (7.25)$$

But by squaring we have introduced a spurious set of solutions. We have satisfied (7.23b) but only one set of solutions gives the cosines in (7.23a) with the correct signs. They are

$$\zeta = \frac{3}{2(\beta^3-3\beta^2+4)}\left[m\pi - (-1)^m \arcsin\left(\frac{\frac{1}{2} + \cos^2\xi}{\cos\xi}\sqrt{\frac{\beta}{2+\beta}}\right)\right] \qquad (7.26a)$$

$$\text{and} \quad \zeta = \frac{3}{4\beta^3}\left[n\pi + (-1)^n \arcsin\left(\frac{2(\beta\cos^2\xi - 1)}{\sqrt{4 - \beta^2}}\right)\right] \qquad (7.26b)$$

where m and n are integers such that m + n is odd if $\cos\xi > 0$, and even if $\cos\xi < 0$, and we require $\zeta \geqslant 0$. Note that (7.26a) diverges at $\beta = 2$ and (7.26b) diverges at $\beta = 0$. The condition that the arcsines exist is the same for both, viz.

$$\beta \leqslant 8\cos^2\xi/(1 + 4\cos^4\xi) \tag{7.27}$$

Therefore null lines only exist in the hatched region under the graph in fig.7.15. In particular, we see that they may only intersect the plane $W_2 = 0$ if $\beta \leqslant 8/5 = 1.6$, a result first found by Berry. Berry's method for finding the nulls in the plane $W_2 = 0$ was to plot the set of curves defined by (7.26) by hand. There are two classes of curves: m even, n odd and m odd, n even. The intersections of all the curves within each class give nulls. I programmed the computer to do precisely the same thing accurately, and the nulls found are shown on fig.7.16. The caustic given by $\beta = 0$ and $\beta = 2$, and the critical curve $\beta = 8/5$ are also plotted accurately. Note that the W_3 scale is much larger than the W_1 scale in all W_1W_3 plots.

The two classes represent the nulls at either end of the loops which intersect $W_2 = 0$, and this pairing is indicated on fig.7.16. The nulls nearest to $\beta = 8/5$ are unpaired, because once the null line has crossed this critical curve it cannot intersect $W_2 = 0$ again. We see that the central region near the W_3 axis is as we predicted previously, and the change of scale with W_3 is clearly visible. The distorted lattice is indicated by dashed lines, and we see that the "null planes" are indeed curved as Nye suggested.

To check this stationary phase approximation, the exact integral for $E(W_1,0,W_3)$ was computed. The amplitude is shown in fig.7.17a, but the phase plot in fig.7.17b most clearly shows the nulls. Those inside the caustic agree exactly, as far as one can tell by superposing the plots, with fig.7.16, although fig.7.17b becomes rather ambiguous near $\beta = 8/5$.

Finally we shall solve eqns.(7.26) for the full three-dimensional configuration of the null lines near $W_2 = 0$. We plot the W_1W_3 and W_1W_2 projections of a particular set of null lines in fig.7.18. This confirms all our expectations. The loops have much the same shape everywhere, but on a smaller scale as W_3 increases. Along a close-packed direction (i.e. the righthand set on fig.7.18) the sides of the loops get closer together as W_3 increases, until outside $\beta = 8/5$ the loops link up into

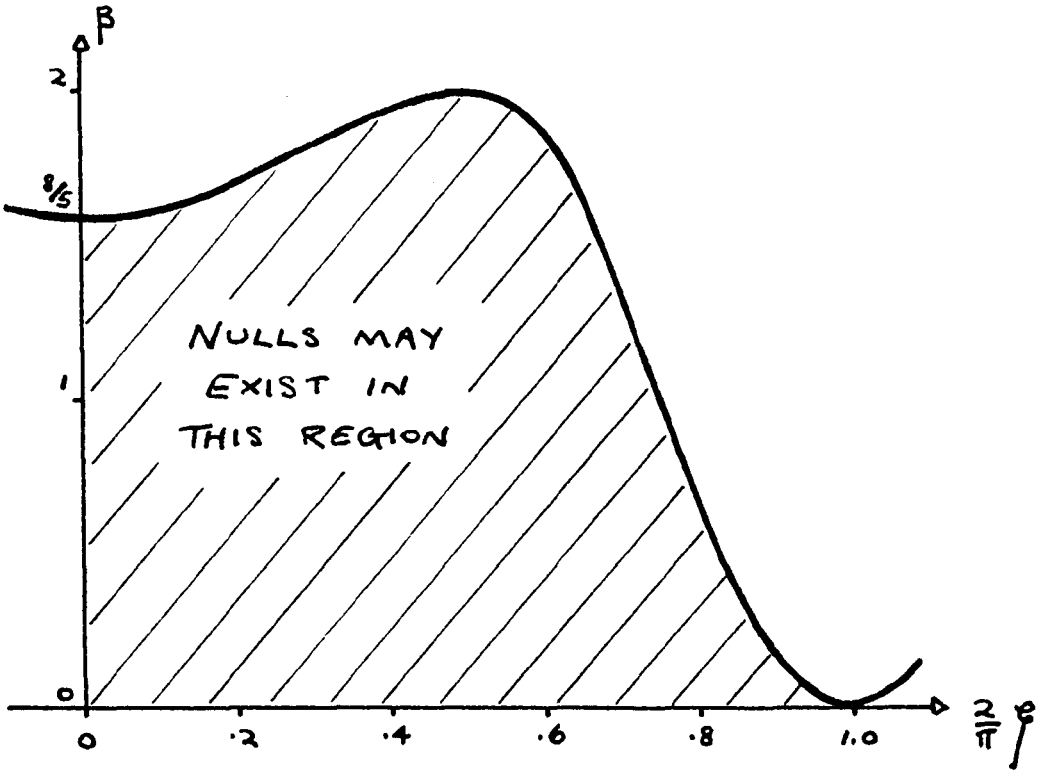
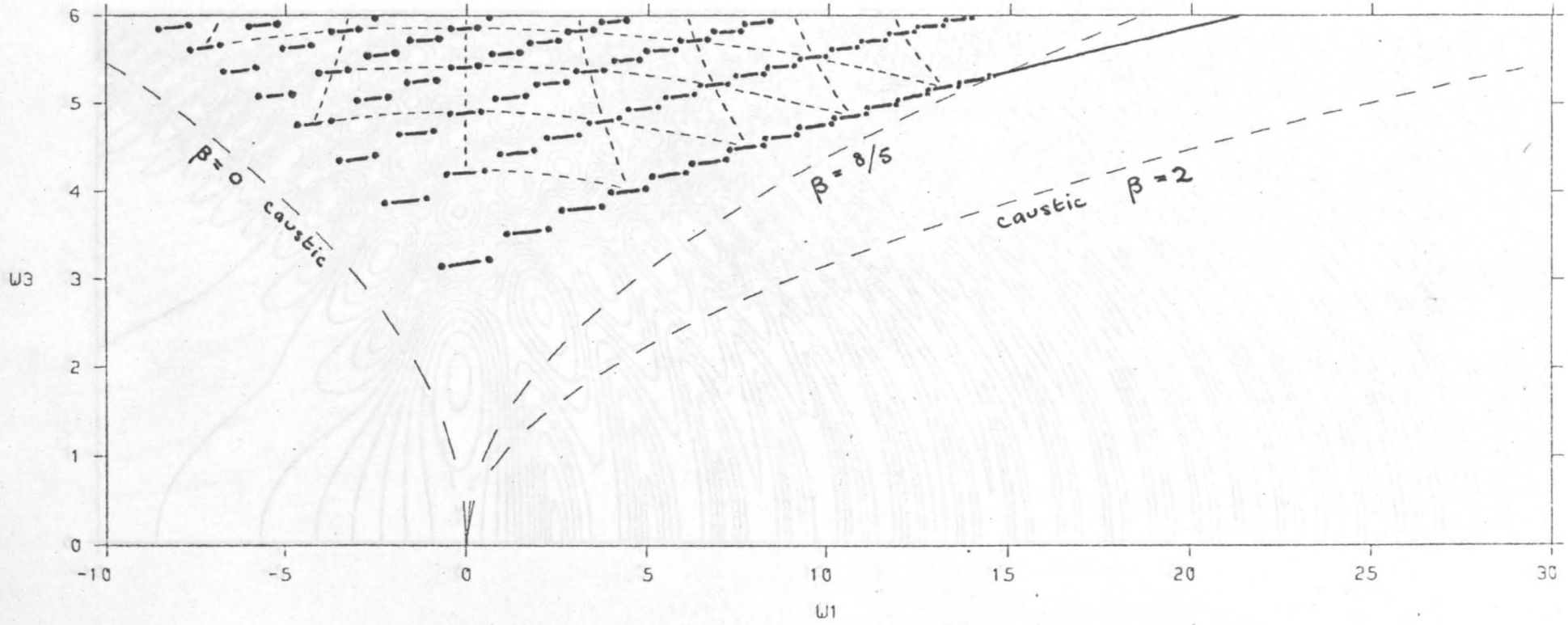
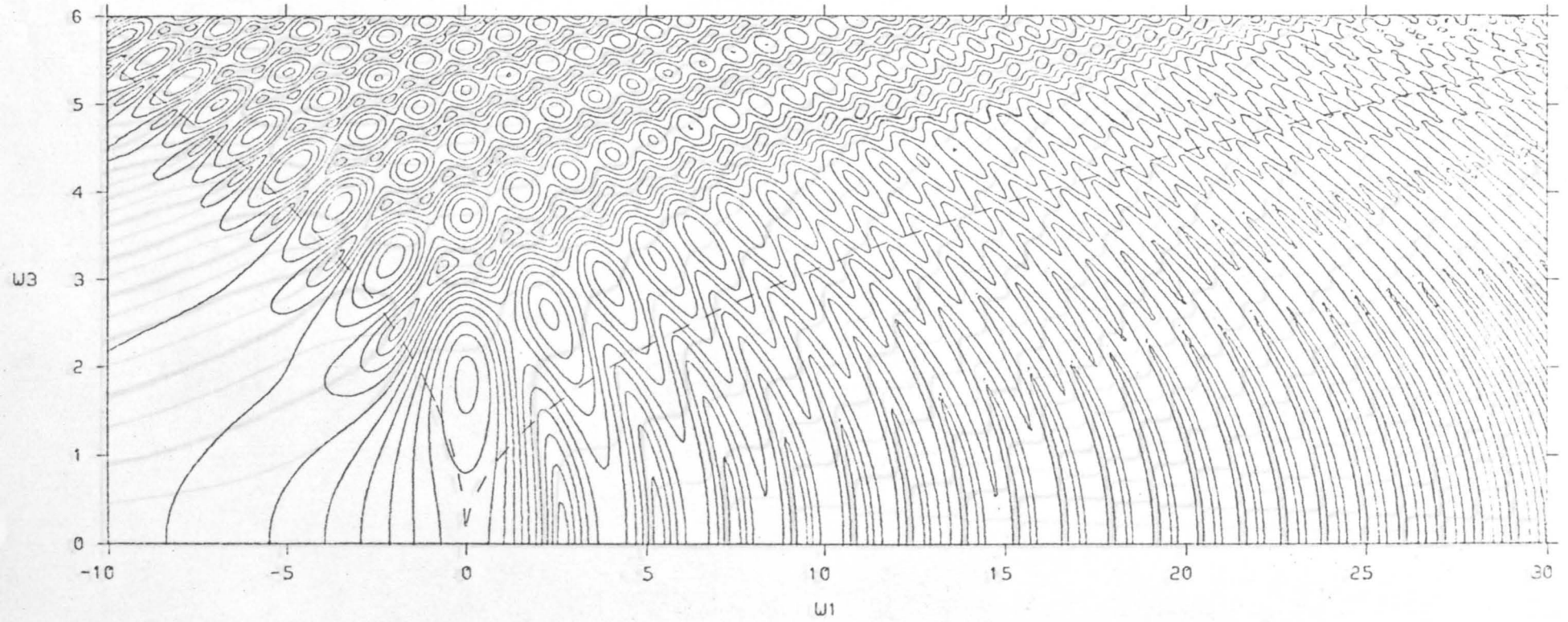


Figure 7.15 Allowed Region for Null Lines



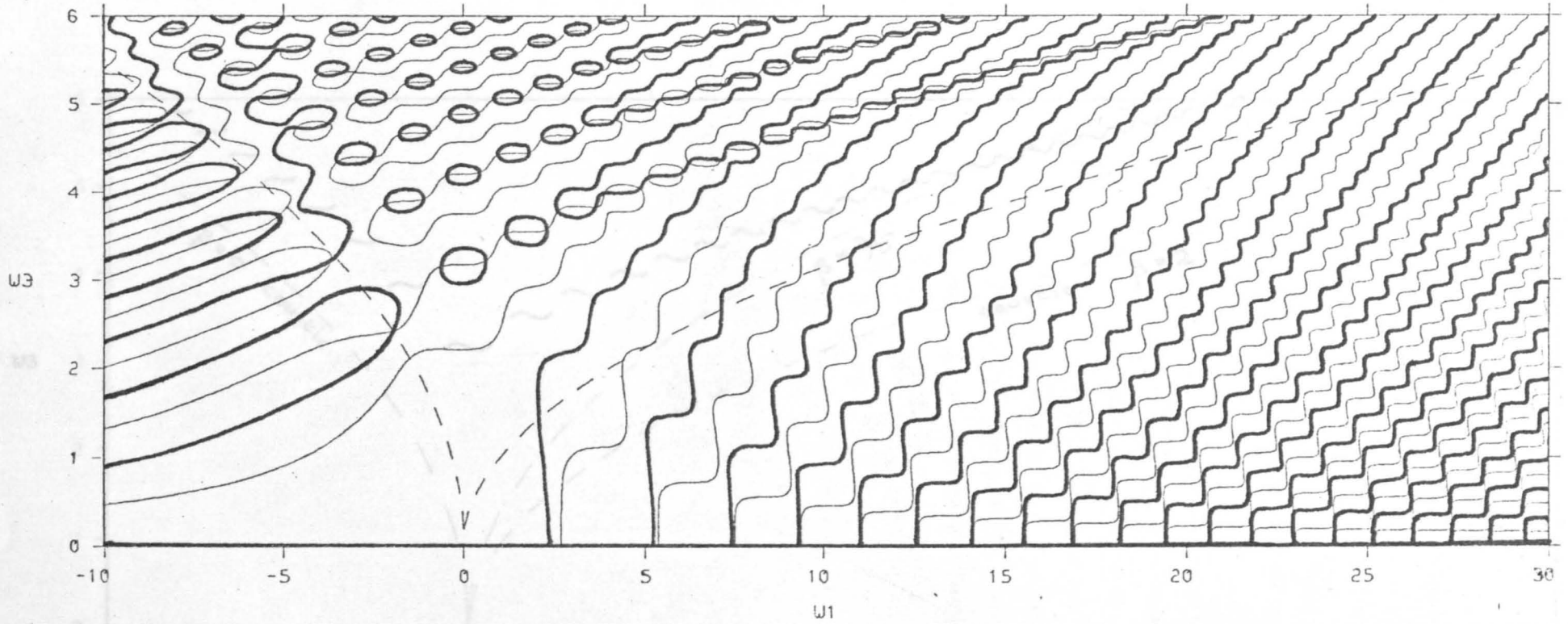
ELLIPTIC UMBILIC [W2=0 SECTION]
NULL POINTS by 'exact' stationary phase showing pairing, and distorted lattice (dashed)

Figure 7.16



ELLIPTIC UMBILIC [W2=0 SECTION]
CONTOURS OF AMPLITUDE in steps of 0·05

Figure 7.17a

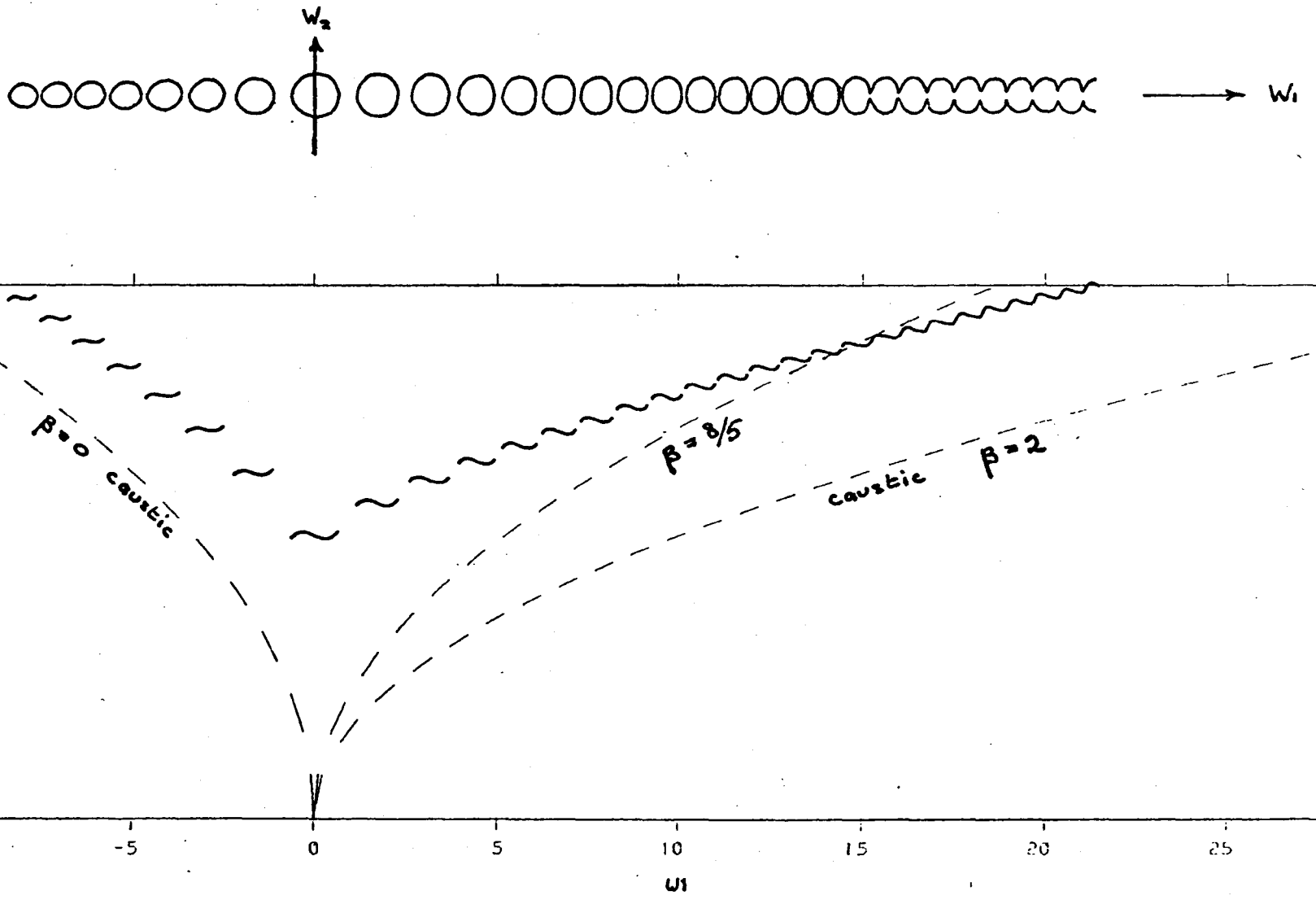


ELLIPTIC UMBILIC [W2=0 SECTION]
CONTOURS OF PHASE in steps of Pi/2

Thick lines 0,π

Thin lines ±π/2

Figure 7.17b



ELLIPTIC UMBILIC [W1-W3 PROJECTION]
NULL LINES by 'approx' stationary phase

Figure 7.18

exactly the hairpin with helical arms postulated by Nye.

The method of solving (7.26) is the following piecewise computation. Fig.7.15 shows that solutions with opposite signs of $\cos\xi$ are quite separate. For a particular choice of sign for $\cos\xi$ the arcsines can only have a certain maximum range of values, so that the values of ζ given by equations (7.26) lie within bands. These intersect to give an array of lozenges labelled by (m,n), within which the (β,ζ) values of a particular segment of null line must lie. We can compute the correct solutions of the cubics in β, resulting from equating the two ζ's at opposite ends of each lozenge, to find the β range encompassed. For each of a number of equally spaced β values in the lozenge, we find the range of ξ allowed by (7.27), and split this into a number of equally spaced values. We have now set up a lattice, at every point of which we can evaluate eqns.(7.26). Then we compute the contour along which the two ζ's are equal, and at every coordinate on this contour we interpolate the actual value of ζ. It then only remains to map these (β,ξ,ζ) values back into $\underline{W}$ space.

This technique could be applied to rows of nulls further away from $W_2 = 0$, and for any values of W_1W_3 inside the caustic. Then by using the symmetry of E, it would be possible to build a fairly complete picture of the nulls inside the caustic.

APPENDIX

A7.1 Computational Details

To evaluate (7.15) we must compute two integrals of the form

$$\int_0^\infty dx\ e^{i\phi_\pm(x)}$$

where the phase $\phi_\pm(x)$ may diverge at both limits of integration, as x^6 as $x \to \infty$ and as $1/x^2$ as $x \to 0$. However, it can easily be shown that the integral converges. The main contributions to the integrals come from the neighbourhoods of the SPPs given, from (7.15), by

$$\chi_\pm(x) = W_2{}^2/6x^3$$

where $\chi_\pm(x) = 6x^5 \mp 8W_3x^3 - 2(W_1-W_3{}^2)x$.

We need only consider $x, W_1, W_2, W_3 \geqslant 0$. The behaviour of $\chi_\pm(x)$ and $W_2{}^2/6x^3$ is sketched in fig.A7.1. The SPPs are given by the intersections of the two curves. For $W_2 > 0$, ϕ_- always has 1 SPP, and ϕ_+ may have 1 or 3 SPPs. As W_1 or W_2 increase the SPPs tend to move to larger x, and as the caustic is crossed from inside, the pair of SPPs of ϕ_+ at smaller x coalesce and disappear. For $W_2 = 0$ the behaviour is slightly different, because both ϕ_+ and ϕ_- have a SPP at $x = 0$ if $W_1 \leqslant W_3{}^2$. But because the SPP is at the lower limit of the integrals, it only contributes half its normal value, and the total number of SPPs is conserved. When $W_3 = 0$, $\phi_+ = \phi_-$ and the two integrals are identical. Otherwise

$$\phi_{1+}(x) - \phi_{1-}(x) = -16W_3x^3 \leqslant 0$$

$$\Rightarrow \quad \phi_{1+}(x) \leqslant \phi_{1-}(x)$$

so the SPP of ϕ_- is always at smaller x than any SPP of ϕ_+.

We define cutoffs $A_\pm$, $B_\pm$ to lie above and below respectively all SPPs of $\phi_\pm$, and we define $G_\pm(x)$ by an obvious extension of equation (4.6). Then $G_\pm(\infty) = 0$ and $G_\pm(0) = 0$ unless $W_2 = 0$, when $G_\pm(0)$ diverges. But for $W_2 = 0$ the phase does not diverge at the lower limit, therefore

$$\int_0^\infty dx\ e^{i\phi_\pm(x)} = G_\pm(A_\pm) + \int_0^{A_\pm} dx\ e^{i\phi_\pm(x)} \quad \text{if } W_2 = 0$$

$$G_\pm(A_\pm) - G_\pm(B_\pm) + \int_{B_\pm}^{A_\pm} dx\ e^{i\phi_\pm(x)} \quad \text{if } W_2 \neq 0$$

At $x = 0$, we only ever compute $\phi_\pm(x)$, and then only when $W_2 = 0$. We avoid the indeterminacy of $W_2{}^2/12x^2$ by replacing it by

$$\lim_{\varepsilon \to 0} \frac{W_2{}^2}{(\varepsilon+12x^2)} .$$

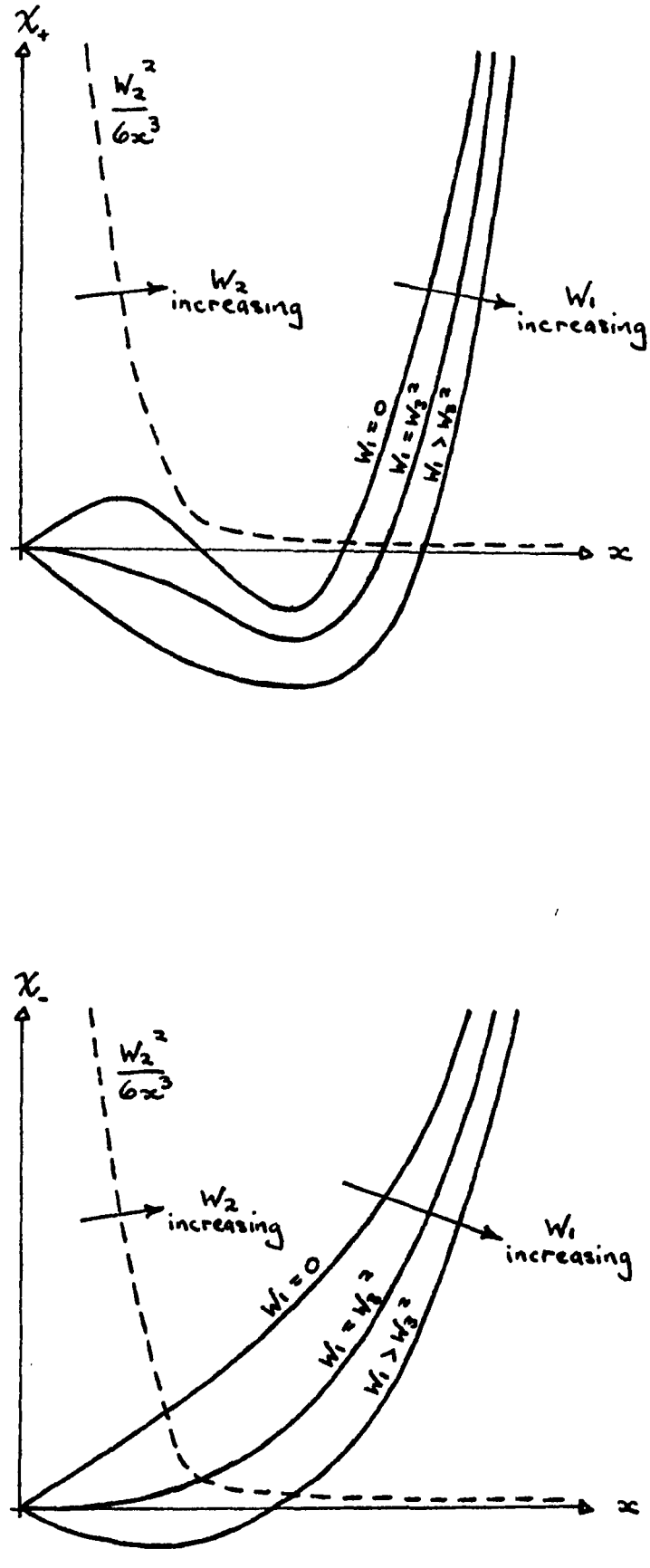


Figure A7.1 Behaviour of $\phi_{\pm}(x)$

Computationally we set ε to a negligibly small value to avoid a divide-by-zero error.

As W_3 increases, the SPPs of ϕ_+ move to larger x, and those of ϕ_- move to smaller x. If we tried to use the same cutoffs for both integrals their separation would be unnecessarily large leading to inefficient quadrature and a likelihood of spurious convergence. Therefore, it is imperative to treat the two integrals quite independently. To plot W_1W_2 sections, our stepping pattern will be to increase W_2 from 0, and for each W_2 to increase W_1 from 0 to its maximum. We find the initial A cutoffs by solving $\phi_{1\pm}(x) = C1$ using Newton's iteration, starting from a large value to ensure that we are outside the SPPs at largest x. We then step out for convergence, and step $A_\pm$ out progressively as we increase W_1. At the start of each W_1 scan we save $A_\pm$ to use as starting values for the next scan.

When W_2 is very small, fig.A7.1 shows that $B_\pm$ must be very small. Then we must examine the convergence of G(x) for small x. From chapter 4, if the general term of G(x) is a_n, then

$$|a_n| = \left|\left(\frac{1}{\phi_1}\frac{d}{dx}\right)^n \frac{1}{\phi_1}\right| \quad \text{and} \quad \phi_1 \sim -\frac{W_2{}^2}{6x^3}$$

$$\Rightarrow \quad |a_n| \sim \frac{(2n+1)!}{2^n\, n!} \times \left(\frac{6x^2}{W_2{}^2}\right)^{n+1} \quad \Rightarrow \quad \left|\frac{a_{n+1}}{a_n}\right| \sim (2n+3)\frac{6x^2}{W_2{}^2} .$$

G(x) is clearly an asymptotic series for small x as well as for large x. This is because, as explained in chapter 4, the method of producing the series G(B) is equivalent to expanding $1/\phi_1(\phi)$ in a Taylor series about $\phi(B)$. As $x \to 0$, $\phi \sim W_2{}^2/12x^2$ and

$$\frac{1}{\phi_1(\phi)} \sim \frac{-W_2}{4\sqrt{3}}\left(\frac{1}{\phi}\right)^{3/2}$$

which clearly has a branch point at the lower limit of integration, where $1/\phi = 0$. Therefore we have integrated up to the circle of convergence of the series in the integrand, so there is no guarantee that the resulting series will converge, and in fact it does not.

In practice, as for G(A), this does not matter as long as the first three terms of G(B) which we use are converging. If we require $|a_{n+1}/a_n| < 1$ for $n = 3$, this implies $x < 0.14\ W_2$. For $W_2 = 0$ we set $B = 0$ and G(B) is not used. As W_1 and W_2 increase, B increases. We start the first $W_2 \neq 0$ scan with $B = 0.1\ W_2$ and decrease B if necessary for convergence. If B becomes ≤ 0 we reduce the step-back. If it is

not necessary to decrease B, we attempt to increase it while retaining convergence. This algorithm should maintain B optimal and allow it to move up towards A, after the pair of SPPs of ϕ_+ disappear on crossing the caustic. At the start of each W_1 scan we save the values of B to initiate the next scan.

The asymptotic forms for large x in chapter 4 give the magnitudes of the first three terms of G(A) as

0.053, 0.0078 & 0.0025

for a series error of less than 0.0025. We use $|\phi_1| > 20$, and require the ratio of the second to first terms of G(x) < 0.25 to ensure that G is actually converging, rather than just small. We use these figures for all four series, so that the total series error should be less than 0.01.

To compute the section $W_2 = 0$, we take B = 0 and do not use G(B). $A_\pm$ both increase with increasing W_1. Therefore we scan with W_1 increasing to push the cutoffs out as usual. But to find the initial cutoffs for each scan we solve $\phi_{1\pm}(A_\pm) = C1$ using Newton's iteration, starting from the values saved from the start of the previous scan.

In order to provide the contour routine with enough data points to be able to handle the fine detail of the diffraction pattern, data at points half-way between those computed were calculated by a cubic 4 point Lagrange interpolation. This must be applied to the real and imaginary parts of the wavefunction, since the non-analytic amplitude and phase cannot be interpolated everywhere.

It was found impossible to get a standard quadrature routine to converge for the elliptic umbilic, and therefore a special routine, ideally suited to this type of integral was developed from a method used by Dr.R.Saktreger, to whom I am most grateful. Saktreger's method is to approximate the phase of the integrand by a linear function in each subinterval, instead of approximating the real or imaginary part by a polynomial, since the phase is varying much more slowly. We use

$$\int_a^b dx\, e^{i\phi(x)} = \frac{h}{2}\int_{-1}^{+1} dy\, e^{i(\sigma + \tau y)} = h\, e^{i\sigma}\, \frac{\sin \tau}{\tau}$$

where $h = b - a$, $\sigma = \frac{1}{2}\{\phi(b) + \phi(a)\}$ and $\tau = \frac{1}{2}\{\phi(b) - \phi(a)\}$. By expanding about $x = \frac{1}{2}(a + b)$ we can show that the magnitude of the error in the linear approximation is

$$\frac{2}{3}\left(\frac{h}{2}\right)^3 \left|\phi_2\left(\frac{a+b}{2}\right)\right| + O(h^4).$$

To estimate the error in the subintegral we compute

$$I_h = \int_a^{a+h} dx\ e^{i\phi} \qquad \text{and} \qquad I_{h/2} = \int_a^{a+h/2} dx\ e^{i\phi} + \int_{a+h/2}^{a+h} dx\ e^{i\phi}$$

and take the error to be

$$\varepsilon_h = |I_h - I_{h/2}|$$

But $\varepsilon_h \propto h^3$, so the subinterval h_{new} necessary to give an error ε_o is given by

$$h_{new} = h(\varepsilon_o/\varepsilon_h)^{1/3} \qquad \text{(A7.2.1)}$$

We start at a = A with a small enough value of h, compute I_h, $I_{h/2}$, ε_h and h_{new}. If $\varepsilon_h > 2\varepsilon_o$ we repeat the process using h_{new}, otherwise we take $I_{h/2}$ as the value of the subintegral and compute the next subintegral with a = A + h, h = h_{new} etc. This method of estimating the required subinterval, as used by Saktreger, is known as Richardson extrapolation. The error in each subintegral is typically ε_o, and certainly less than $2\varepsilon_o$. Empirically, increasing ε_o by a factor of 10 reduces the typical step length by a factor of 2, therefore the actual error using $I_{h/2}$ is probably only about 0.1 times the estimated error. A value for ε_o of 0.0005 was found satisfactory, giving a maximum total error of about 0.01 for a typical number of 20 subintervals. ε_o must be chosen small enough to avoid spurious convergence.

To plot the caustic dashed in the positive quadrant, we use the fact that

$$\frac{ds}{d\theta} = \pm\frac{4}{3} W_3^2 \sin\frac{3\theta}{2}$$

from (7.5), so we find successive points from

$$\theta_{n+1} = \theta_n + \frac{\delta s}{\frac{4}{3} W_3^2 \sin\frac{3\theta_n}{2}}$$

This diverges at $\theta_0 = 0$, but putting $\theta = d\theta$ gives

$$\theta_1 = (\delta s/2W_3^2)^{\frac{1}{2}}$$

and we proceed from there.

CHAPTER 8

WAVEFRONT DISLOCATIONS AS CATASTROPHES

8.1 General Theory

We began in chapter 1 by discussing wavefront dislocations using a complex wavefunction as a model, and analysed their behaviour in some special cases. In chapter 4 we showed how caustics in wavefields are catastrophes, and we analysed the continuous wave dislocations (null lines) which arise in the caustic diffraction patterns. I now present a few concluding remarks to show that the dislocations themselves are catastrophes. The advantage of a catastrophe theory of wavefront dislocations is that it only requires a real wavefunction, and the theory is much closer to experiment than the complex wavefunction theory. Also once we can apply the machinery of catastrophe theory we should be able to analyse the structural stability of different types of dislocation, although I only present a preliminary discussion here.

Consider a real wavefunction $\psi(\underline{r},t)$ and observe it at fixed $\underline{r}$ as a function of t. In practice we would do this by placing a receiver at $\underline{r}$, and observing the signal on an oscilloscope. Then $\underline{r}$ is our control variable, and t our state variable. Crests and troughs will occur on the display when

$$\frac{\partial\psi}{\partial t}(\underline{r},t) = 0 \qquad (8.1)$$

and will be crests or troughs as $\partial^2\psi/\partial t^2 < 0$ or > 0. At a particular time t_0, (8.1) defines a surface (of many sheets) in space, which we could plot out experimentally. As t varies, these surfaces will move in space. The family of surfaces at different values of t constitutes the catastrophe manifold, and the crest/trough surface in space at any instant t_0 is the section $t = t_0$ through this manifold. The crest/trough surface will generally consist locally of a set of independent crest and trough surfaces. However, these may join up along a fold, as in fig.8.1. The set of surfaces will sweep upwards with time, so that on moving from point A to point C we lose a crest and a trough. This will occur at point B on the fold line, where

$$\frac{\partial^2\psi}{\partial t^2}(\underline{r},t) = 0$$

This is precisely what one would call the dislocation line experimentally. The "fold line" is actually a 2-surface in $(\underline{r},t)$, analogous to what we

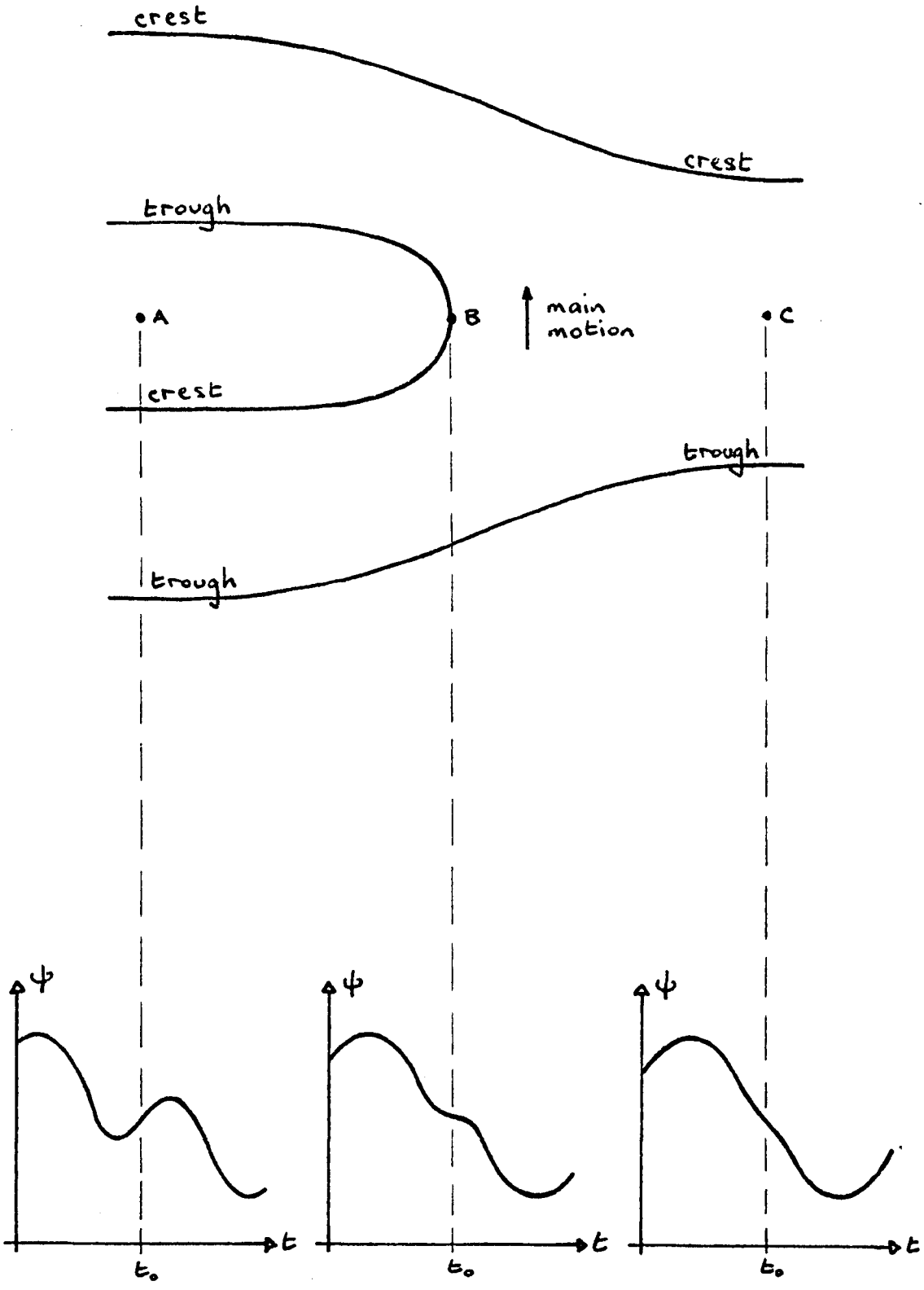


Figure 8.1 Catastrophe Manifold near a Dislocation

previously called the dislocation 4-trajectory. In catastrophe theory the projection of the "fold line" into control space $\underline{r}$ is the bifurcation set, and for a dislocation this is the real space 3-trajectory. Catastrophe theory tells us what local forms of the wavefunction $\psi(\underline{r},t)$ about a dislocation line are structurally stable: in fact, we are restricted to the fold, cusp and swallowtail catastrophes. We shall attempt to interpret these catastrophes as dislocations. If we regard a crest/trough pair as constituting a wavefront, then clearly the fold line represents the end of a wavefront. However, the crest/trough surfaces behave somewhat differently from the equiphase surfaces which are normally regarded as wavefronts (see chapter 1), therefore to avoid confusion we shall rarely use the term wavefront in this chapter.

For simplicity, we shall consider essentially a two-dimensional control space, and we shall ignore the other sheets of the catastrophe manifold around the fold, which give rise to undislocated "wavefronts". Suppose the catastrophe manifold has a simple fold, not lying in a plane parallel to the control plane, which projects into a regular curved fold bifurcation set, as in fig.8.2. Let the top sheet of the manifold represent a crest. The plane $t = t_0$ generally intersects the crest sheet in a curve, ending on the fold line at the dislocation. As t varies, the crest line and dislocation move together. By changing the global shape of the manifold, one can make the crest line swing round the dislocation, or make the dislocation appear to move along the crest line. Glide and climb are motions of the dislocation line relative to the asymptotic wavefronts, and since catastrophe theory is a local theory we can strictly only discuss "local" glide or climb. However, the dislocation is always attached to the same crest line, so it is not very meaningful to talk about glide in these terms, but climb is quite possible.

The speed with which the dislocation moves along its trajectory depends upon the angle the fold line makes locally with the control plane. In the limit where the fold line lies in a plane parallel to the control plane, the dislocation moves along its trajectory with infinite velocity. We shall ignore this degenerate case. The opposite extreme is continuous waves, where the dislocations are fixed in space. Then in the CW limit the trajectories must degenerate into points, which means the fold lines must be straight and parallel to the t axis. Such singularities are not catastrophes, and are therefore not structurally stable. This instability is shown by the fact that if we make a small

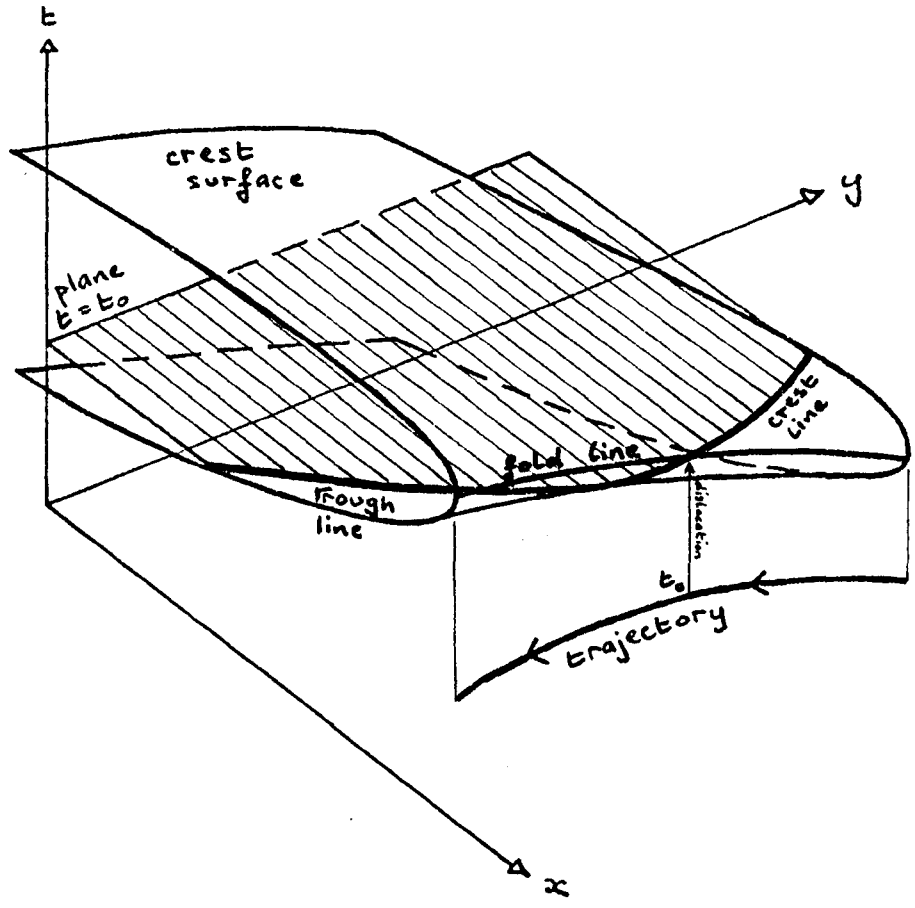


Figure 8.2 Fold Catastrophe as a Dislocation

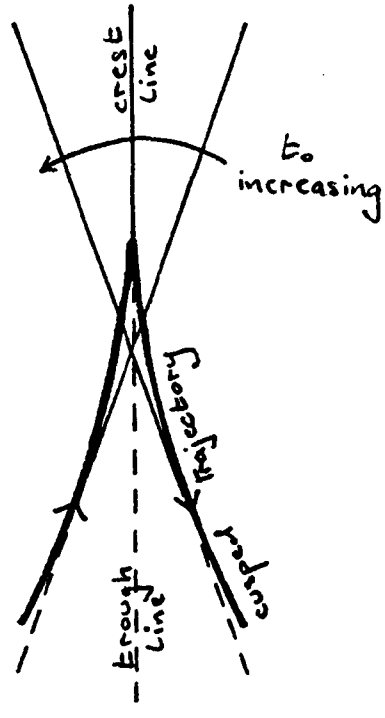


Figure 8.3 Crest/Trough Lines around a Cusped Trajectory

general perturbation to a wave (i.e. do not just change its frequency) its dislocations will change discontinuously from being fixed in space to moving along some trajectory.

We can produce birth/death events by curving the fold line up or down so that the actual times of dislocations along the trajectory are changed, without necessarily changing the trajectory. Hence these events are not significant in catastrophe theory. Also, by considering only a two-dimensional system, we are restricting ourselves to pure edge dislocations. The edge/screw distinction is a global characteristic, not a local one. A screw dislocation line is perpendicular to the asymptotic wavefronts, but locally the dislocation line must always lie in the wavefront surface. We begin to realize that the complex wavefunction theory of dislocations, and the catastrophe theory, find completely different features of a dislocation to be significant. This could be important in the experimental detection and characterization of dislocations.

A static dislocation (see chapter 1) is easy to handle within our catastrophe framework because

$$\psi(x,y,z,t) = \Psi(x,y,t - z/c) \tag{8.2}$$

is essentially two-dimensional. We can explore the catastrophe manifold either by varying t as before, or by varying z at fixed t. Therefore, alternate surfaces of the catastrophe manifold represent the actual crest surfaces in 3-space at some fixed time, and the fold line is actually the dislocation line. This structure sweeps rigidly up the z axis as time increases, and the dislocation line sweeps out a trajectory surface which is the bifurcation set, as before. On this definition, a structurally stable static dislocation line can have any shape allowed by catastrophe theory for the fold line. This is a regular curve, and singularities will only be apparent in the trajectory surface, as we shall see in the next section.

The next catastrophe is the cusp, whose standard catastrophe manifold is illustrated in fig.4.3. The plane $t = t_0$ intersects this manifold in straight lines. As t_0 increases these straight lines "roll round" the cusped bifurcation set B, as shown in fig.8.3. Suppose the catastrophe manifold outside B represents a crest: then the "middle sheet" inside B between the two branches of the fold line represents a trough. So the crest line in fig.8.3 becomes a trough line (shown dashed) where it touches B, and the point where it touches is the dislocation.

At the cusp point the dislocation stops its motion and doubles back on itself. For a non-standard cusp the straight lines would become curved, but the essential behaviour must be the same.

If we incorporate z into t as in (8.2) to give a static dislocation, then the whole catastrophe manifold minus the "middle sheet" represents a crest surface sweeping up the z axis as time increases. Asymptotically the manifold is parallel to the xy (control) plane, and at the cusp catastrophe the fold line is normal to this plane. Therefore at this point the dislocation line is perpendicular to the asymptotic crest surface, which is what we call a screw dislocation. Then for static dislocations it seems reasonable to say that a dislocation at a cusp catastrophe point has pure screw character. On this definition a dislocation can only have pure screw character in a structurally stable way at isolated points! Another significance of the cusp catastrophe, which is always true, is that the wavefunction is locally purely symmetric in time, so that the crest and trough appear or disappear symmetrically (see fig.2 of NB74).

With our model having only two essential control variables, we have shown that motion, birth and death of dislocations can occur in a structurally stable manner. Because this is a local theory, distinctions between glide and climb, edge and screw, can not really be made, since they are defined relative to the asymptotic wavefronts. Locally, all dislocations are of edge type, and they can only climb (relative to a crest surface, rather than a wavefront, which is not strictly defined). The nearest we have got to a screw dislocation is the cusp catastrophe point in a static dislocation, which can only occur at isolated points.

The remaining structurally stable behaviour involves the swallowtail catastrophe, whose manifold exists in 4 dimensions. The dislocation line produced by the standard form of the catastrophe will be a straight line in three-dimensions, and the crest surfaces will be planes rolling round the bifurcation set. Since the swallowtail manifold has 4 sheets inside the tail, we have the possibility of two dislocations interacting. If additional parameters are included in the wavefunction we have the possibility of higher catastrophes, but the standard forms of these will always give a straight dislocation line and plane crest surfaces. Rather than pursue these ideas further, we shall investigate the relationship between the two theories of dislocations by analysing in terms of catastrophe theory the simplest dislocation discussed in chapter 1.

8.2 The Canonical Single Strength Dislocation

In section 1.3 we showed that the complex wavefunction in the neighbourhood of a static straight single strength dislocation of mixed screw-edge type is

$$\psi(x,y,z,t) = \{x + \beta_s y + \beta_e(z-ct)\}\, e^{ik(z-ct)}$$

where β_s and β_e are not both real. To apply catastrophe theory we require a real wavefunction, therefore to preserve generality we shall include a constant phase ϕ before taking the real part. We define $T = \omega t - kz$ and factor out β_e for convenience (assuming $\beta_e \neq 0$), and take as our general real wavefunction

$$\psi = \mathrm{Re}(\alpha x + \beta y + T)e^{i(\phi - T)}$$

$$= (X + T)\cos(T - \phi) + Y\sin(T - \phi)$$

where $X \equiv \alpha_R x + \beta_R y$, $Y \equiv \alpha_I x + \beta_I y$ are more convenient coordinates. If we shift T and X by $T \to T + \phi$, $X \to X - \phi$ we have

$$\psi = (X + T)\cos T + Y\sin T \tag{8.3}$$

Thus by diffeomorphic transformations of our variables we have absorbed all the constants into modified space and time variables. In particular, which real "projection" of our complex wavefunction we take, has only the effect of shifting the origin in the XT plane.

Now we regard $\psi(T,X,Y)$ as a potential function and look for its bifurcation set. Solving

$$\partial\psi/\partial T = \partial^2\psi/\partial T^2 = 0 \tag{8.4}$$

gives

$$X = \pi/2 \pm \{\arctan\sqrt{-(Y+2)/(Y+1)} - \sqrt{-(Y+2)(Y+1)}\} \tag{8.5}$$

Solving $\partial^3\psi/\partial T^3 = 0$ with (8.4) shows that cusp points occur at

$$X = -T = -(n + \tfrac{1}{2})\pi, \quad Y = -2.$$

The bifurcation set, sketched in fig.8.4, is a z = constant section through the trajectory surface. It looks like a projection of a helix, and in fact the fold line is a distorted helix. If we define

$$\xi = X + T, \quad \eta = Y + 3/2$$

then the solution of (8.4) can be written as

$$\xi^2 + \eta^2 = \tfrac{1}{2}^2, \text{ a circle of radius } \tfrac{1}{2}, \text{ and } \tan T = -(\tfrac{1}{2} - \eta)/\xi.$$

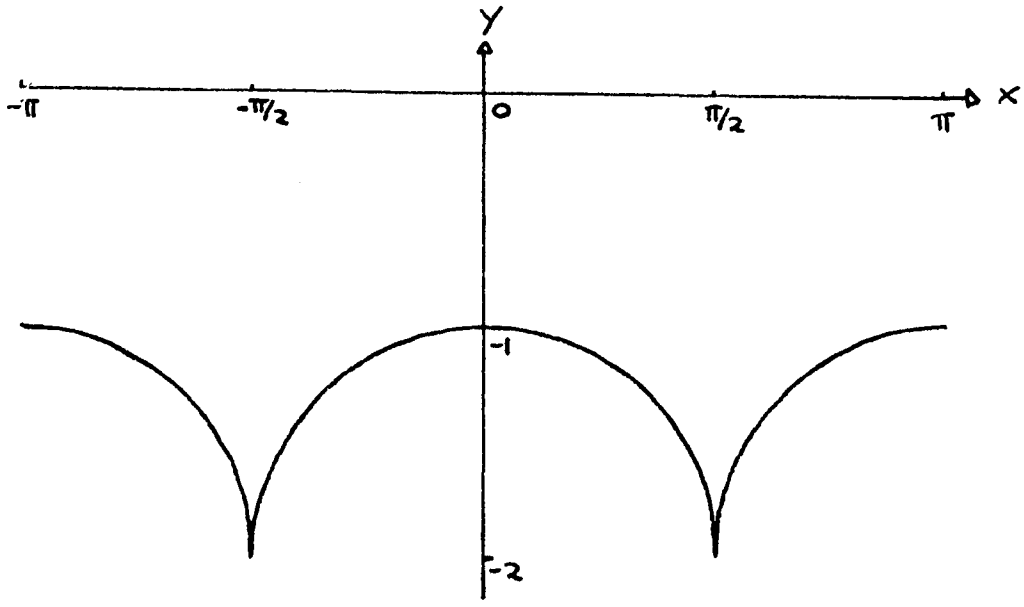


Figure 8.4 Bifurcation Set for Canonical Single Straight Dislocation

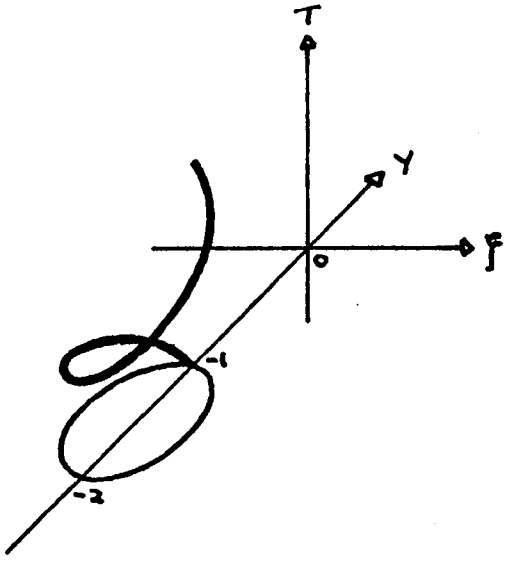


Figure 8.5a Helical Fold Line in (ξ,Y,T) Coordinates

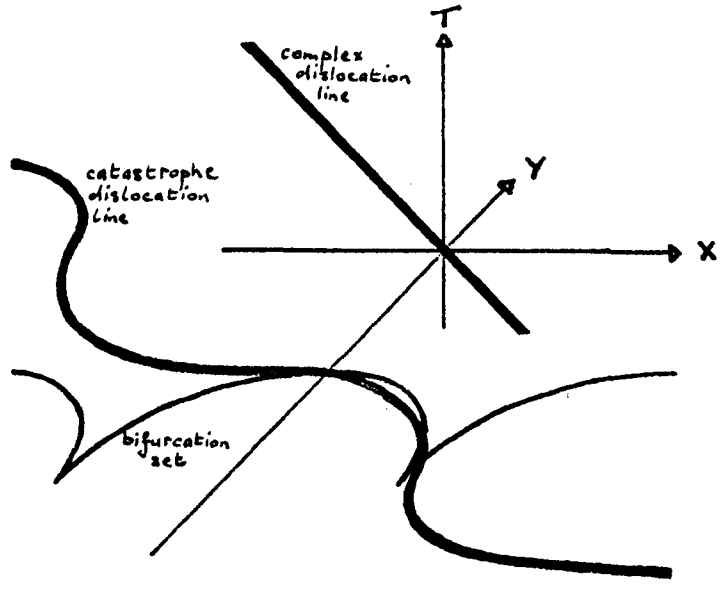


Figure 8.5b Dislocation Lines and Bifurcation Set

Elementary geometry shows that T is the distance around the circumference of the projection onto the (ξ,η) plane. Then the fold line in (ξ,η,T) coordinates is a regular circular helix of radius $\frac{1}{2}$ with a pitch of 45^0, as sketched in fig.8.5a. To map this back into (X,Y,T) space we must shear it along ξ by the transformation $X = \xi - T$, so that the axis of the helix now lies at 45^0 to the (X,Y) plane. Then the points at $Y = -1$ become parallel to the control plane, as in the standard fold catastrophe, and the points at $Y = -2$ become perpendicular to the control plane, and project into the cusps of the bifurcation set.

At $t = 0$, the catastrophe dislocation line is essentially the fold line, since $T = -kz + \phi$. As t varies, this dislocation line moves rigidly along the z axis, sweeping out the trajectory surface. But the original complex wavefunction in our final coordinates is

$$(X + iY + T)\, e^{-iT}$$

whose "complex" dislocation line (zero of amplitude) is given by

$$X = -T, \quad Y = 0$$

as shown on fig.8.5b.

In the pure edge limit of the complex dislocation $\alpha_R = \beta_R = 0$, and the place of X in (8.3) is taken by ϕ, which is then not removeable. In the pure screw limit of the complex dislocation $|\alpha|$ and $|\beta| \to \infty$ so that the scale of the mapping of xy into XY increases. In xy space the catastrophe dislocation line approaches the T axis and in the limit coalesces with it. Then the fold line is a straight line perpendicular to control space, as we found earlier for CW dislocations, and this is no longer a catastrophe. Therefore, the pure screw dislocation is not structurally stable. The pure edge dislocation is structurally stable in space-time except for values of ϕ which put us on a YT section through a cusp. In this case the dislocation will be symmetric, so a symmetric pure edge dislocation is not structurally stable in space-time alone. Any other pure edge dislocation is structurally stable, as is the whole family of pure edge dislocations with different values of ϕ.

It is very difficult to perform the above analysis for any more complicated model dislocated wavefunction, therefore the way to proceed further is probably from the "catastrophe end", as in section 1, rather than from the "dislocation end". We have shown that there is a close link between dislocations and catastrophes, which is worthy of further study.

ACKNOWLEDGEMENTS

It is a pleasure to thank my supervisor, Dr. Mike Berry, for his boundless but undemanding enthusiasm, and for supplying me with ideas (at a rate somewhat faster than I could implement them), many of which form the basis of this dissertation. I would also like to thank Prof. J.F.Nye, Victor Humphrey and many other colleagues for many valuable discussions. Their ideas have become too closely interwoven with my own for me to be able to give them adequate credit other than in this general way.

I would specifically like to thank Dr. Raymond Saktreger for explaining his numerical integration method to me; George Keene for doing all the magnificent photography necessary to produce the simulated diffraction patterns in chapter 7; Dianne Cokayne for allowing me to use her excellent IBM Selectric typewriter (without its interchangeable typeheads and error-correction ribbon the production of this dissertation would have been considerably more arduous than it was!); and Colin Upstill for his fastidious proof-reading.

All computing was done on the ICL System 4 computers in the South West Universities Computer Network, and I am most grateful to the staff of the Computer Centre for their assistance. I thank the Science Research Council for providing this computer time, and also for supporting me on a Research Studentship.

REFERENCES

Abramowitz M. & Stegun I.A. (Eds.) 1965 Handbook of Mathematical Functions. Dover

Airy G.B. 1838 Proc. Camb. Phil. Soc. 6, 379-402

Bateman Manuscript Project (Ed. A. Erdelyi) 1954 Tables of Integral Transforms. McGraw Hill

Beaver W.L. 1974 J. Acoust. Soc. Am. 56, 1043

Bell W.W. 1968 Special Functions for Scientists and Engineers. Van Nostrand

Berry M.V. 1972 J. Phys. A5, 272-291

Berry M.V. 1975 J. Phys. A8, 566-84

Berry M.V. 1976 Adv. Phys. 25, 1-26

Berry M.V. & Nye. J.F. 1977 Nature 267, 34-6

Berry M.V., Nye. J.F. & Wright. F.J. 1978 in preparation

Bjorken J.D. & Drell S.D. 1964 Relativistic Quantum Mechanics. McGraw Hill

Bleaney B.I. & Bleaney B. 1965 Electricity and Magnetism. OUP

Bröcker Th. 1975 Differential Germs and Catastrophes. CUP

Budden K.G. 1961 Radio Waves in the Ionosphere. CUP

Carter A.H. & Williams A.O. 1951 J. Acoust. Soc. Am. 23, 179

Chester C., Friedman B. & Ursell F. 1957 Proc. Camb. Phil. Soc. 53, 599-611

Connor J.N.L. 1973 Faraday Disc. Chem. Soc. No.55, 51-7
and Mol. Phys. 26, 1371-7

Connor J.N.L. 1976 Mol. Phys. 31, 33-55

Coulson C.A. 1965 Waves. Oliver & Boyd

Dennery P. & Krzywicki A. 1967 Mathematics for Physicists. Harper International

Dingle R.B. 1973 Asymptotic Expansions: their derivation and interpretation. London: Academic Press

Dirac P.A.M. 1931 Proc. Roy. Soc. A133, 60

Erdelyi A. 1956 Asymptotic Expansions. New York: Dover

Farn C.L.S. & Huang H. 1968 J. Acoust. Soc. Am. 43, 252

Flett T.M. 1966 Mathematical Analysis. London: McGraw Hill

Freedman A. 1970 J. Acoust. Soc. Am. 48, 135

Hilbert D. & Cohn-Vossen S. 1952 Geometry and the Imagination. New York: Chelsea Publishing Co.

HCV52 ditto

Hirschfelder J.O. et al. 1974 J.Chem. Phys. 61, 5435-59

Hirschfelder J.O. et al. 1976 J.Chem. Phys. 64, 760-85

Hirschfelder J.O. et al. 1976 J.Chem. Phys. 65, 470-86

Huntington, Emslie & Hughes 1948 J. Franklin Inst. 245, 1

Jackson J.D. 1962 Classical Electrodynamics. Wiley

Landau L.D. & Lifschitz E.M. 1975 The Classical Theory of Fields. Pergamon

Levine H. & Schwinger J. 1948 Phys. Rev. 74, 958

Liesegang S. 1953 Optik 10, 5-14

Lighthill M.J. 1958 Introduction to Fourier Analysis and Generalized Functions. CUP

Lockwood J.C. & Willette J.G. 1973 J. Acoust. Soc. Am. 53, 735

Marini & Rivenez 1974 Ultrasonics 12, 251

Maslin N.M. 1976 J. Atmos. Terr. Phys.

Mather J.N. 1972 Right Equivalence. University of Warwick

Morse P.M. 1948 Vibration and Sound. McGraw Hill

Nye J.F. 1978 Proc. Roy. Soc. A in press

Nye J.F. & Berry M.V. 1974 Proc. Roy. Soc. Lond. A336, 165-90

NB74 ditto

Oberhettinger F. 1961 J. Res. Nat. Bur. Stand. (US) B65, 1

Pearcey T. 1946 Phil. Mag. 37, 311-7

Poston T. & Stewart I.N. 1976 Taylor Expansions and Catastrophes. London: Pitman

Rayleigh, Lord 1878 Theory of Sound. Vol.II, Chapter XI. MacMillan

Riess J. 1970 Ann. Phys. 57, 301-21

Riess J. 1970 Phys. Rev. D2, 647-53

Riess J. 1976 Phys. Rev. B13, 3862-9

Rindler W. 1960 Special Relativity. Oliver & Boyd

Rschevkin S.N. 1963 Theory of Sound. Pergamon

Rutherford D.E. 1965 Vector Methods. Oliver & Boyd

Schoch A. 1941 Akust. Z. 6, 318

Spence R.D. 1949 J. Acoust. Soc. Am. 21, 98

Stepanishen P.R. 1971 J. Acoust. Soc. Am. 49, 283

Stepanishen P.R. 1971 J. Acoust. Soc. Am. 49, 841

Stepanishen P.R. 1971 J. Acoust. Soc. Am. 49, 1629

Tables of Integral Transforms, Bateman Manuscript Project (Ed. A. Erdelyi) 1954. McGraw Hill

Thom R. 1975 Structural Stability and Morphogenesis. Reading, Mass: Benjamin

Trinkaus H. & Drepper F. 1977 J. Phys. A10, L10-6

Walford M.E.R., Holdorf P.C. & Oakberg R.G. 1977 J. Glaciol. 18, 217-29

Wassermann G. 1974 Stability of Unfoldings. Springer Math. Notes 393

Watson G.N. 1944 A Treatise on the Theory of Bessel Functions. Edn.2. CUP

Williams A.O. 1951 J. Acoust. Soc. Am. 23, 1

Woodcock A.E.R. & Poston T. 1974 A Geometrical Study of the Elementary Catastrophes. Springer Math. Notes 373

Zemanek J. 1971 J. Acoust. Soc. Am. 49, 181